Karolina Grzech

Analiza Danych Jakościowych - Projekt zaliczeniowy
RAPORT

Tematem badania będzie określenie wpływu różnych czynników na jakość czerwonej oraz białej odmiany portugalskiego wina “Vinho Verde”, której miarą jest ich końcowa ocena. Bazę analizy stanowią dane znajdujące się pod adresem:

https://www.kaggle.com/datasets/shelvigarg/wine-quality-dataset/data

Zbiór zawiera 6497 obserwacji na temat jakości wina oraz zawartych w nim składników. Oprócz wyników jakości każdego z win, plik zawiera informacje na temat składu każdego z nich. Celem badania będzie analiza wpływu czynników na wyniki jakosci. Aby rozważane dane były bardziej „zwarte” (to znaczy żebyśmy nie mieli wartosci z każdą możliwą zawartoscią poszczególnego skladnika oraz ocen wina w skali 0-10), dokonamy dyskretyzacji zmiennych oraz zmiennej objasnianej. Wino uznawane będzie za dobre jakościowo, jeżeli jego ocena będzie równa co najmniej 6. Prowadzi to oczywiście do utraty znacznej części informacji o dokładnych ocenach win. Oceny z wynikiem 6/10 oraz 10/10 świadczą o sporej różnicy w smaku oraz jakosci, mimo że w badaniu trafią one do tej samej grupy. Przy takim podejściu możemy jednak stosować metody rozumowania w modelu dwumianowym, gdzie oczywiście za sukces przyjmujemy uznanie wina za dobre. Zmienne ilościowe zostały przemianowane na jakościowe, zgodnie z normami wartości każdego ze składników, w celu znalezienia potencjalnych zależności pomiędzy nimi, a zmienna objaśnianą.

- type
  - red
  - white
- fixed.acidity (Zalecana kwasowość, odpowiednia dla prawidłowej fermentacji mieści się w granicach 6-9 g kwasu/l, stąd też wartości poniżej tego przedziału będziemy traktować jako niskie, a wartości wyższe niż 9, jako wysokie)
- volatile.acidity (Wartość 0.7 uznawana jest za wysoki poziom lotnej kwasowości)
- citric.acid (Na potrzeby analizy dzielimy wartości citric.acid na 5 grup)
  - 5 : >= 1
  - 4 : [0.7,1)
  - 3 : [0.55,0.7)
  - 2 : [0.3,0.55)
  - 1 : < 0.3

- residual.sugar - przemianowana na zmienną jakościową, zgodnie ze standardami poziomu słodkości wina:
    - Sweet : residual.sugar większe od 45
    - Semi-sweet : residual.sugar w przedziale (12,45]
    - Semi-dry : residual.sugar w przedziale (4,12]
    - Dry : residual.sugar mniejsze od 4
- chlorides (Na potrzeby analizy dzielimy wartości chlorides na 6 grup)
    - 6 : >= 0.1
    - 5 : [0.069,0.1)
    - 4 : [0.056,0.069)
    - 3 : [0.048,0.056)
    - 2 : [0.03,0.048)
    - 1 : < 0.03
- free.sulfur.dioxide (Zawartość wolnego i ogólnego dwutlenku siarki w winach jest zróżnicowana, jednakże w żadnym przypadku nie przekracza dopuszczalnej według obowiązujących norm i przepisów zawartości tego konserwantu w winie, odpowiednio, 50 i 260 mg/dm_3) Stąd:
    - High : >= 30
    - Medium : [15,30)
    - Low : < 15
- total.sulfur.dioxide
    - High : >= 200
    - Medium : [100,200)
    - Low < 100
- density (wina o gęstości w granicach 0.97-0.99 uznawane są za musujące)
- pH
    - High : pH większe bądź równe 3.8
    - Recommended : pH w przedziale [3.3;3.8)
    - Medium : pH w przedziale [3.0;3.3)
    - Low : pH mniejsze od 3.0
- sulphates (Na potrzeby analizy dzielimy wartości sulphates na 6 grup)
    - 6 : >= 1.5
    - 5 : [1.0,1.5)
    - 4 : [0.5,1.0)
    - 3 : [0.4,0.5)
    - 2 : [0.3,0.4)
    - 1 : < 0.3
- alcohol (Przeciętna ilość alkoholu w winie to 10%-12% obj.)
    - High : alcohol > 12
    - Medium : 10 < alcohol < 12
    - Low : alcohol > 10

Przyjrzyjmy się najpierw zależności oceny jakosci od rodzaju wina. Poniższa tabela zawiera informacje o winach uznanych za dobre bądź nie, w zależności od rodzaju, wraz z licznościami brzegowymi.

	1 - good	**0 - bad**	
white	**2743**	**1262**	4005
red	**844**	**733**	1577
	3587	1995	

Dokonamy estymacji prawdopodobieństwa uzyskania dobrej oceny przez każdy z typów wina korzystając z podejść Walda, Agrestiego-Coulla, Wilsona i dokładnego. Dla wina białego otrzymujemy metodą największej wiarogodności $\tilde{p}$ = 0,685, zaś metodą przedziałową w każdym podejściu, po zaokrągleniu do dwóch miejsc po przecinku, przedział (0,67; 0,7). W przypadku wina czerwonego, estymatory mają postać $\tilde{p}$ = 0,535 oraz przedziały (0,51; 0,6). Na tej podstawie wysnuwamy hipotezę, iż wino białe cieszy się większym uznaniem. Argumentem potwierdzającym tę hipotezę jest estymator ilorazu szans na poziomie θ’ = 1.89, co interpretujemy tak, że szansa otrzymania zadowalajacej oceny w przypadku wina białego jest znacznie wieksza niż w przypadku wina czerwonego.

Na podstawie wykonanych w R: testu Chi-kwadrat oraz obliczenia współczynnika Goodmana-Kruskala τ = 0,019, przy bardzo niskiej p-wartości, w przybliżeniu równej 0, hipotezę o niezależności wyników od rodzaju wina odrzucamy.

Następnym aspektem naszej analizy będzie zbadanie wpływu ilosci zawartego alkoholu na wyniki. Tabela oraz wyniki estymacji wygladaja nastepujaco

	1	0
High	874	58
Medium	1868	757
Low	845	1180

	$\tilde{p}$	Przedziały
High	0.94	(0.92;0.95)
Medium	0.71	(0.69;0.72)
Low	0.41	(0.4;0.44)

Widać, że im zawartosc alkoholu w winie maleje, szansa na otrzymanie zadowalajacej oceny również jest mniejsze. W przypadku win o najwyzszej zawartosci alkoholu szansa jest dosyć duża, bo w okolicach 90%, gdzie dla win o najmniejszej ilosci jest to jedynie ok 40%.

Współczynnik Goodmana-Kruskala oraz wykonany test Chi-kwadrat również odrzucają hipotezę o niezależności ocen od zawartości alkoholu.

W taki sam sposób przeanalizowałam, czy istnieje wpływ poszczególnych zmiennych na zmienną określającą jakość wina. Analogicznie wykonałam wszystkie tabele, a następnie przeprowadziłam testy. Jedynie w przypadku zmiennej określającej pH nie było podstaw do odrzucenia hipotezy zerowej (tj. odrzucenie niezależności zmiennej pH od jakość wina) .

	1	0
High	8	1
Recommended	1194	621
Medium	2116	1235
Low	269	138

Faktycznie, analizując przedstawione niżej wartości, nie widać zależności związanej ze zmianą oceny jakości pod wpływem zmiany pH. Wykonane testy to potwierdzają.

	$\tilde{p}$	Wald	Agresti	Wilson	dokladne
High	0.88/0.78	(0.68;1.09)	(0.54;0.99)	(0.57;0.99)	(0.52;0.99)
Recommended	0.66	(0.64;0.7)	(0.64;0.7)	(0.64;0.7)	(0.64;0.7)
Medium	0.63	(0.61;0.65)	(0.61;0.65)	(0.61;0.65)	(0.61;0.65)
Low	0.66	(0.64;0.67)	(0.64;0.67)	(0.64;0.67)	(0.64;0.67)

Otrzymana p-wartość, w tym przypadku równa 0.088, nie daje podstaw do odrzucenia hipotezy o niezależności.

W pozostałych analizach hipotezy o niezależności zmiennych zostały odrzucone.

Jednak może się np. okazać, że to nie jeden konkretny składnik znacząco wpływa na ocenę wina, co pewnie skorelowane z nim zmienne dają obserwowany efekt. Możemy zauważyć, że np. korelacja między zmiennymi total.sulfur.dioxide oraz free.sulfur.dioxide wynosi 0.72.

Przeszłam więc do przeanalizowania modelu regresji logistycznej, początkowo uwzględniając wszystkie zmienne, stopniowo zmniejszając ich liczbę, jednocześnie sprawdzając jej statystyki oraz wyniki anovy.

Podsumowanie modelu uwzględniającego wszystkie zmienne wygląda następująco:

```
Call:
glm(formula = good ~ ., family = binomial, data = wine)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           1.203e+02  4.776e+01   2.519 0.011754 *
typewhite            -6.794e-01  1.996e-01  -3.403 0.000666 ***
fixed.acidity         9.281e-02  5.472e-02   1.696 0.089840 .
volatile.acidity     -4.536e+00  3.169e-01 -14.316  < 2e-16 ***
citric.acid          -5.128e-01  2.925e-01  -1.753 0.079588 .
residual.sugar        1.176e-01  2.064e-02   5.697 1.22e-08 ***
chlorides            -2.107e+00  1.167e+00  -1.804 0.071156 .
free.sulfur.dioxide   3.427e-02  3.826e-03   8.956  < 2e-16 ***
total.sulfur.dioxide -7.539e-03  1.179e-03  -6.396 1.60e-10 ***
density              -1.312e+02  4.850e+01  -2.706 0.006816 **
pH                    6.549e-01  3.249e-01   2.016 0.043847 *
sulphates             2.181e+00  2.933e-01   7.436 1.03e-13 ***
alcohol               8.246e-01  6.414e-02  12.857  < 2e-16 ***
```

Jak widać, nie wszystkie zmienne użyte w modelu są istotne. Stopniowo zmniejszałam ilość branych pod uwagę zmiennych i obserwowałam, jak zmieniają się statystyki modelu. Po przekształceniach otrzymałam model uwzględniający: volatile.acidity, residual.sugar, free.sulfur.dioxide, total.sulfur.dioxide, sulphates oraz alcohol.

Jednak po przeanalizowaniu wyników, wartości dewiancji oraz wykorzystując kryterium Akaike(AIC), model uwzględniający interakcje pomiędzy wyżej wymienionymi zmiennymi okazał się najlepszy i to on finalnie zostaje wybrany.

```
>model7 <- glm(good ~ volatile.acidity*residual.sugar*
          free.sulfur.dioxide*total.sulfur.dioxide*
          sulphates*alcohol,family = binomial, data = wine)
```

Wykonane testy (test wiarygodności oraz test Chi-kwadrat) potwierdziły poprawny dobór zmiennych do modelu.

Na koniec sprawdziłam dokładność wybranego przeze mnie modelu. Tablica, przedstawiająca wyniki ma następującą postać:

	Prawdziwe	
Przewidziane	0	1
0	364	152
1	207	952

Dokładność modelu wynosi około 79% i jest to zadowalający wynik.