Klaudia Szczygieł

Projekt zaliczeniowy

Analiza danych jakościowych

1. Wprowadzenie - Opis danych, źródło danych oraz sposób gromadzenia.

Dane poddane analizie pochodzą z repozytorium znajdującego się na platformie kaggle. Opis wraz z danymi pochodzi z następującej strony: https://www.kaggle.com/mohansacharya/graduate-admissions.
Dane te orginalnie zostały zainspirowane innymi danymi, znanymi pod nazwą „UCLA Graduate Dataset" i zostały zmodyfikowane przez Mohan S Acharya. Do analizy zostały wybrane z powodu występowania binarnej zmiennej objaśniającej (akceptacją bądz odrzucenia kandydatury na studia wyższe), pełnej informacji danych (każda jednostka posiada wszystkie cechy). Ramka danych składa się z 9 zmiennych objaśniających i zmiennej objaśnianej. Rekordy zostały zebrane na podstawie badania wykonanego przy udziale 400 osób. Projekt będzie miał na celu zbadanie uogólnionego modelu regresji, konkretnie wpływu poszczególnych zmiennych które pozwolą sklasyfikować na jakiej podstawie rozpatrzeć kandydaturą osoby ubiegającej się o dostanie się na studia magisterskie.

- **GRE** - wynik egzaminu dyplomowego, zmienna ilościowa
- **TOEFL score** - wynik testu językowego, potwierdzający znajomość języka obcego na danym poziomie. Zmienna ilościowa
- **University Rating -** ranking uniwersytetu (zmienne o wartościach uporządkowanych, 5 poziomów: 1 - najgorszy, 5 - najlepszy)
- **SOP** (statement of purpose) - wynik listu motywacyjnego. Zmienne o wartościach uporządkowanych (9 poziomów: od 1 do 5)
- **LOR (Letter of Recommendation)** - wynik listu rekomendacyjnego. Zmienne o wartościach uporządkowanych (9 poziomów: od 1 do 5)
- **CGPA** - średnia z Collegu. Zmienna ilościowa
- **Research** - przeprowadzone badania naukowe. Zmienna ilościowa binarna (0 - kandydat nie przeprowadził badania, 1 - kandydat przeprowadził badanie)
- **Get admission** - objaśniana zmienna binarna (0 - kandydat nie przyjęty, 1 - kandydat został przyjęty)

2. **Analiza danych.**

Wstępna analiza danych

Pierwszym krokiem będzie wykonie wstępnej analizy danych dla każdej zmiennej. Podstawowe informację dotyczące każdej z danych takich jak na przykład mediana lub średnia w próbie, jak również liczności dla zmiennych jakościowych prezentują się następująco:

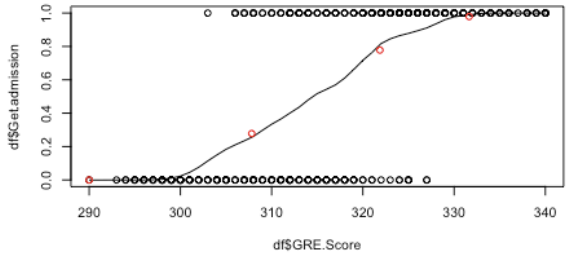
```
   GRE.Score      TOEFL.Score    University.Rating      SOP
 Min.   :290.0   Min.   : 92.0   Min.   :1.000     Min.   :1.0
 1st Qu.:308.0   1st Qu.:103.0   1st Qu.:2.000     1st Qu.:2.5
 Median :317.0   Median :107.0   Median :3.000     Median :3.5
 Mean   :316.8   Mean   :107.4   Mean   :3.087     Mean   :3.4
 3rd Qu.:325.0   3rd Qu.:112.0   3rd Qu.:4.000     3rd Qu.:4.0
 Max.   :340.0   Max.   :120.0   Max.   :5.000     Max.   :5.0
```

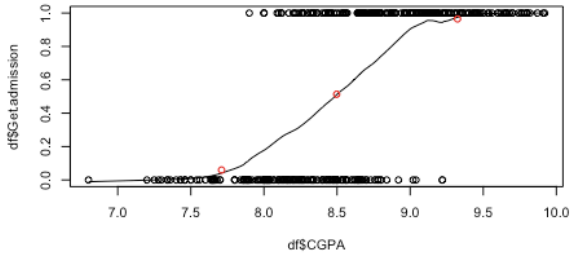
```
      LOR             CGPA           Research      Chance.of.Admit  Get.admission
 Min.   :1.000   Min.   :6.800   Min.   :0.0000   Min.   :0.3400   Min.   :0.000
 1st Qu.:3.000   1st Qu.:8.170   1st Qu.:0.0000   1st Qu.:0.6400   1st Qu.:0.000
 Median :3.500   Median :8.610   Median :1.0000   Median :0.7300   Median :1.000
 Mean   :3.453   Mean   :8.599   Mean   :0.5475   Mean   :0.7244   Mean   :0.565
 3rd Qu.:4.000   3rd Qu.:9.062   3rd Qu.:1.0000   3rd Qu.:0.8300   3rd Qu.:1.000
 Max.   :5.000   Max.   :9.920   Max.   :1.0000   Max.   :0.9700   Max.   :1.000
```

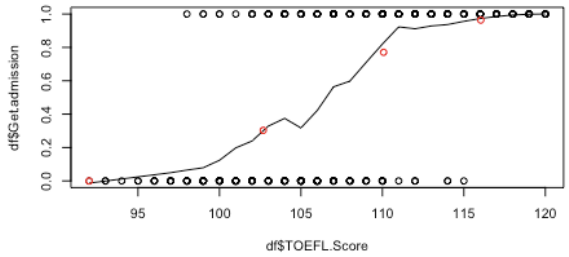
W badaniu wzięło udział 400 osób, w tym 196 osób którym nie udało się zaklasyfikować na studia magisterskie oraz 204 osoby których kandydatura zakończyła się sukcesem. Średni wynik z egzaminu dyplomowego wyniósł 316.8, natomiast średni wynik z testu z języka obcego wyniósł 107.4.

Analiza zależności pomiędzy dwoma zmiennymi

Wykresy zależności zmiennej objaśnianej od zmiennych ilościowych







W przypadku zmiennych GRE.Score, TOEFL.Score, CGPA można obserwować tendencję wzrostową do ekstremalnych wartości

Przejdziemy teraz do sprawdzenia zależności pomiędzy zmienną objaśnianą a zmiennymi objaśniającymi jakościowymi. W tym celu wykorzystamy testy badające zależności pomiędzy zmiennymi jakościowymi.

Jako pierwszą zmienną sprawdzimy związek pomiędzy przeprowadzanymi badaniami (Research) a przyjęciem na uczelnie. Poniżej zostaje przedstawiona tabela kontyngencji.

Tabela 1

	NIEPRZYJĘTY	**PRZYJĘTY**
NIEZROBIONE BADANIA	128	53
ZROBIONE BADANIA	46	173

Po przeprowadzeniu testu zależności zmiennych Research oraz przyjęcie na uczelnie w oparciu o współczynnik τ Goodmana - Kruskala uzyskano wartość statystyki τ = 0.2491246 oraz p-wartość na poziomie równym 2.062489e-23. Obliczenia sugerują, że zrobienia badań ma istotny wpływ na dostanie się na uczelnie wyższą (ponieważ jeżeli p < 0.05 to można podejrzewać powiązanie zmiennych, im tau bliższe 1 tym też lepsze dopasowanie). Z powyższej tabeli wnioskujemy, że osoby które zrobiły badania zostają częściej przyjęte. Na podstawie analizy ilorazów szans uzyskujemy estymator na poziomie 0.1100975 oraz p-wartość = 1.305251e-21. Przedział ufnośći na poziomie 0.95 wynosi (0.06976064 ; 0.17375799). Analiza ilorazów szans wskazuje na zależność zmiennych. P-wartość dla testu χ^2 oraz Fishera jest mniejsza niż 2.2e-16, co również odrzuca nam hipotezę o niezależności zmiennej.

Kolejną zmienną jaką będziemy testować ze zmienną objaśnianą, to ocena uniwersytetu.

Tabela 2

OCENA UNIWERSYTETU	NIEPRZYJĘTY	PRZYJĘTY
1	25	1
2	80	27
3	56	77
4	9	65
5	4	56

Po przeprowadzeniu testu zależności powyższych zmiennych na podstawie współczynnika Goodmana-Kruksala uzyskujemy wartość τ = 0.1261718 oraz p-wartość na poziomie 3.062535e-10. Natomiast dla testu χ^2 p-wartosć jest równa 2.882e-10 oraz dla testu Fishera otrzymaliśmy p-wartość = 4.466e-11.Testy Goodmana-Kruksala, χ^2 oraz Fishera odrzuciły hipotezę o niezależności.

W następnym kroku zbadamy związek pomiędzy zmienną objaśnioną a wynikiem listu motywacyjnego oraz listu rekomendacyjnego.

Tabela 3

OCENA LISTU MOTYWACYJNEGO	NIIEPRZYJĘTY	PRZYJĘTY
1	6	0
1.5	20	0
2	27	6
2.5	31	16
3	34	30
3.5	33	37
4	16	54
4.5	5	48
5	2	35

Dla powyższej tablicy współczynnik Goodmana-Kruksala wynosi 0.1585788 oraz p-wartość = 1.058419e-10 więc hipotezę o niezależności odrzucamy. Mała liczność pierwszego wiersza nie pozwala nam na użycie asymptotyki testu χ^2.

Tabela 4

OCENA LISTU REKOMENDACYJNEGO	NIEPRZYJĘTY	PRZYJĘTY
1	1	0
1.5	5	2
2	35	3
2.5	24	15
3	55	30
3.5	32	41
4	18	59
4.5	3	42
5	1	34

Współczynnik Goodmana-Kruksala wyniósł tutaj 0.2298896 oraz p-wartość = 3.165523e-19, wartosci te nie dają nam powodów aby odrzucić hipotezę o niezależności. Podobnie jak w przypadku analizowania Tabeli 3 liczność pierwszego wiersza nie pozwala nam na użycie asymptotyki testu χ^2.

3. Specyfikacja i weryfikacja modelu regresyjnego.

W tej części przejdziemy do dopasowania modelu regresji logistycznej przedstawiającej wpływ czynników na dostanie się na studia magisterskie. Modele będziemy porównywać ze sobą obliczając dla nich wartości tzw. Kryterium informacyjne Akaike (AIC). Im mniejsza wartość kryterium AIC, tym lepiej dopasowany model. W pierwszej kolejności podzielimy danych na zbiór w którym wykonamy estymację modelu (75%) oraz oraz na zbiór walidacyjny (25%), gdzie będziemy badać zdolności prognostyczne modelu.
W drugim kroku sprawdziliśmy GLM dla zmiennej objaśnianej wraz z uwzględnieniem wszystkich zmiennych objaśniających.

Ogólny model regresyjny z logitową funkcją wiążącą:

```
Call:
glm(formula = y_train ~ ., family = binomial(link = logit), data = X_train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.29651  -0.42940   0.07746   0.43719   2.19532

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -56.77114    9.40878  -6.034  1.6e-09 ***
GRE.Score           0.12365    0.03732   3.314 0.000921 ***
TOEFL.Score         0.02614    0.06188   0.422 0.672788
University.Rating   0.20173    0.26958   0.748 0.454263
SOP                 0.16595    0.31409   0.528 0.597260
LOR                 0.95272    0.33456   2.848 0.004404 **
CGPA                1.23847    0.69062   1.793 0.072929 .
Research            0.39238    0.41091   0.955 0.339625
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 411.06  on 299  degrees of freedom
Residual deviance: 187.68  on 292  degrees of freedom
AIC: 203.68

Number of Fisher Scoring iterations: 6
```

Ogólny model regresyjny z logitową funkcją wiążącą:

```
Model 1: y_train ~ GRE.Score + TOEFL.Score + University.Rating + SOP +
    LOR + CGPA + Research
Model 2: y_train ~ 1
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       292     187.68
2       299     411.06 -7  -223.38 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Z powyższych danych wynika, że model uwzględniający wszystkie zmienne objaśniające jest niewystarczający, ponieważ wiele zmiennych jest nieistotnych statystycznie (brak podstaw do odrzucenia hipotezy o zerowaniu się współczynników w modelu). Test Walda odrzucił hipotezę zerową o nieistotności jedynie dla zmiennych GRE.Socre, LOR (na poziomie ufności 95%). Te zmienne uwzględnimy w naszym końcowym modelu. Ponadto analiza testu wiarygodności modelu pełnego do modelu z wyrazem wolnym wykazała wyższość modelu pełnego do modelu z wyrazem wolnym wykazała wyższość modelu pełnego (p-wartość < 2.2e-16)

W kolejnym kroku przystąpimy do redukcji zmiennych. Modele będzie porównywać za pomocą kryterium Akaike.

```
                 Step Df  Deviance Resid. Df Resid. Dev      AIC
1                    NA        NA       292   187.6848 203.6848
2       - TOEFL.Score  1 0.1783323       293   187.8632 201.8632
3               - SOP  1 0.3673758       294   188.2306 200.2306
4          - Research  1 0.9781549       295   189.2087 199.2087
5 - University.Rating  1 1.3759664       296   190.5847 198.5847
```

Po jego wykonaniu modelem z najniższym współczynnikiem (AIC = 198.6) jest model bez zmiennych TOEFL.Score, SOP, Research, University.Rating.

```
Call:
glm(formula = y_train ~ . - TOEFL.Score - SOP - Research - University.Rating,
    family = binomial(link = logit), data = X_train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.24428  -0.44429   0.08096   0.42659   2.34646

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -63.0157     8.4571  -7.451 9.25e-14 ***
GRE.Score     0.1466     0.0311   4.714 2.43e-06 ***
LOR           1.1108     0.2974   3.735 0.000188 ***
CGPA          1.5413     0.6248   2.467 0.013625 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 411.06  on 299  degrees of freedom
Residual deviance: 190.58  on 296  degrees of freedom
AIC: 198.58

Number of Fisher Scoring iterations: 6
```

Model zredukowany posiada nieznacznie większą dewiancję, jednak jego wartość kryterium AIC jest niższa. Ponadto test istotności Walda wykazał, że każda ze zmiennych w modelu okazała się istotna statystycznie. Wykonaliśmy również test ilorazu wiarygodnosci dla przyjętego modelu względem modelu pełnego. Statystyka testu obliczona jako róznica dewiancji obu modeli zwróciła wartość -2.8998 oraz p-wartość równą 0.5747. Wartości zwrócone przez testy statystyczne sugerują dobre dobranie modelu regresji logistycznej.

```
Model 1: y_train ~ GRE.Score + TOEFL.Score + University.Rating + SOP +
    LOR + CGPA + Research
Model 2: y_train ~ (GRE.Score + TOEFL.Score + University.Rating + SOP +
    LOR + CGPA + Research) - TOEFL.Score - SOP - Research - University.Rating
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       292     187.69
2       296     190.59 -4  -2.8998   0.5747
```

Dodatkowo sprawdzimy, czy nie występuje problem współliniowości przy użyciu funkcji vif(). Eliminacja współliniowości polega na usunięciu z modelu cech, które są kombinacją liniową innych zmiennych niezależnych. Jeżeli wartość współczynnika VIF > 10, wówczas możemy spodziewać się współliniowości. Poniższa tabela przestawia wartość wpółczynników VIF dla poszczególnych zmiennych. Zaobserwowano dość silną współliniowość pomiędzy zmiennymi, co może być problematyczne dla naszego modelu.

```
GRE.Score        LOR       CGPA
 39.71901   20.91752   43.61829
```

W celu detekcji zależności między zmiennymi spójrzmy na korelację między zmiennymi

```
           GRE.Score       LOR      CGPA   y_train
GRE.Score 1.0000000 0.5509559 0.8431710 0.6802225
LOR       0.5509559 1.0000000 0.6782049 0.5724707
CGPA      0.8431710 0.6782049 1.0000000 0.6852798
y_train   0.6802225 0.5724707 0.6852798 1.0000000
```

Walka ze współliniowością w modelu jest trudna. Najprostszym sposobem jest zwiększenie liczby obserwacji w modelu. Kolejną metodą w walce z zależnością pomiędzy zmiennymi jest wyrzucenie zmiennych, które są podejrzewane o współniowość w modelu. Jak widać zmienna CGPA jest silnie skorelowana ze zmienną GRE.Score. Aby pozbyć się współliniosć pomiędzy zmiennymi usuniemy z naszego modelu zmienną CGPA ponieważ jest ona skorelowana ze zmienną GRE.Score, ponadto CGP ma mniejszą korelację ze zmienną objaśnioną. Po wyrzuceniu z naszego modelu zmiennej CGPA AIC wzrosło do 290.57. Niestety usunięcie zmiennej skorelowanej pogorszyło nasz model. W związku z tym najlepszym modelem jest model ze zmiennymi GRE.Score, LOR, CGPA pomimo zaskakująco wysokich wartości VIF.

```
Call:
glm(formula = Get.admission ~ . - TOEFL.Score - SOP - Research -
    University.Rating - CGPA, family = binomial(link = logit),
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3445  -0.5224   0.1216   0.5093   2.2648

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -62.44804    6.79000  -9.197  < 2e-16 ***
GRE.Score     0.18719    0.02136   8.762  < 2e-16 ***
LOR           1.07331    0.20634   5.202 1.97e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 547.74  on 399  degrees of freedom
Residual deviance: 287.28  on 397  degrees of freedom
AIC: 293.28

Number of Fisher Scoring iterations: 6
```

Interpretacja modelu

W tej części przejdziemy do interpretacji współczynników oraz ilorazów szans dla zmiennych objaśniających w przyjętym modelu. Szansa jest to iloraz prawdopodobieństwa sukcesu do porażki. W regresji logistycznej iloraz szans jest to exp(c), gdzie c jest współczynnikiem danej zmiennej.

Współczynniki zmiennych w dopasowanym modelu regresji:

ZMIENNA	**Wyraz wolny**	**GRE.Score**	**LOR**	**CGPA**
WSPÓŁCZYNNIK	-63.0157	0.1466	1.1108	1.5413

Kolejna tabela zawiera szanse oraz przedziały ufności dla oszacowań współczynników.

	ILORAZ SZANS	**2,5%**	**97,5%**
Wyraz wolny	4.291699e-28	-80.93362116	-47.6042411
GRE.Score	1.157891	0.08858257	0.2111862
LOR	3.036787	0.54780036	1.7198015
CGPA	4.670658	0.34027539	2.7986404

Po przyjrzeniu się powyższym ilorazom szans dla poszczególnych zmiennych można wysnuć następujące wnioski. Średnia z Colleagu (CGPA) determinuje w sposób znaczący dostanie się na studia magisterskie. Wynosi one aż 4.67. W przypadku wyniku z egzaminu dyplomowego (GRE.Score) iloraz szans wynosi 1.15, co może sugerować brak większego wpływu na dostanie się na studia. Dobrze napisany list motywacyjny zwiększa szanse pozytywnej kandydatury 3-krotnie

5. Predykcja na zbiorze walidacyjnym.

odel regresji logistycznej może zostać wykorzystany do prognozy szans dostania się kandydata na studia magisterskie spoza wykorzystywanego zbioru danych. Prognoza zostanie wykonana dla 25% danych których nie użyliśmy podczas modelowania.

```
          Reference
Prediction  0  1
         0 40 19
         1  3 38

               Accuracy : 0.78
                 95% CI : (0.6861, 0.8567)
    No Information Rate : 0.57
    P-Value [Acc > NIR] : 8.958e-06

                  Kappa : 0.5708
 Mcnemar's Test P-Value : 0.001384

            Sensitivity : 0.9302
            Specificity : 0.6667
         Pos Pred Value : 0.6780
         Neg Pred Value : 0.9268
             Prevalence : 0.4300
         Detection Rate : 0.4000
   Detection Prevalence : 0.5900
      Balanced Accuracy : 0.7984

       'Positive' Class : 0
```

Na podstawie powyższej danych, widzimy że model który ewaluowaliśmy na podstawie zbioru treningowego działa całkiem dobrze na zbiorze testowym. Dokładność modelu (czyli liczba poprawnie sklasyfikowanych próbek podzielona przez całkowitą liczbę rekordów) wynosi 0.78, co jest całkiem dobrym wynikiem. Wrażliwość modelu (Sensitivity) wynosi 0.93, natomiast specyfikacja (Specificity) modelu 0.67. Z powyższych statystyk wynika również, ze model lepiej przewiduje wartości negatywne (40 próbek negatywnych sklasyfikowanych jako negatywne, tylko 3 próbki negatywnie sklasyfikowane jako

pozytywne), niż pozytywne (19 niepoprawnych klasyfikacji - próbki były pozytywne a dostały etykietę negatywnej oraz 38 próbek sklasyfikowanych poprawnie jako pozytywne).

Wrażliwość oraz specyficzność modelu oblicza się za pomocą poniższych wzorów:

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{38}{38 + 3} = 0.93$$

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{40}{40 + 19} = 0.67$$

gdzie:

- TP (true positive) - próbki które prawidłowo zostały sklasyfikowane jako prawdziwe
- FN (false negative) - próbki które nieprawidłowo zostały sklasyfikowane jako negatywne
- TN (true negative) - próbki które prawidłowo zostały sklasyfikowane jako negatywne
- FP (false positive) - próbki które nieprawidłowo pzostały sklasyfikowane jako pozytywne

6. Wnioski

Podsumowując, analiza danych czterystu osób za pomocą regresji logistycznej umożliwiła wyspecyfikowanie z 8 zmiennych zmiennych najbardziej wpływowych na dostanie się bądz odrzucenie kandydatury na studia magisterskie. Po wykluczeniu nieistotnych zmiennych, ostatecznie model posiada 3 zmienne, z których najbardziej wpływowe okazały się być średnia z Colleagea (prawie 4,5 większe szanse), list rekomendacyjny (dobra opinia zwiększa szansę aż dwukrotnie). Wynik egzaminu dyplomowego wydają się mieć nikły dodatni wpływ na dostanie się na uczelnie, co jest trochę sprzeczne z realiami panującymi w Polsce. Na tą zmienną może mieć wpływ zmienna niosąca informację o średniej z Colleagu (zmienna są ze sobą skorelowane). Odrzucone w analizie zmienne to wynik z egzaminu TOEFL, poziom uniwersytetu z którego pochodzi kandydat oraz wynik listu motywacyjnego. Dopasowany na podstawie kryterium Akaike model logistyczny wydaję się być zgodny z intuicją (list rekomendacyjny jest ważniejszy od samooceny kandydata, średnia z egzaminu dyplomowego wraz ze średnią ze studiów liecencjackich jest istotniejsza niż wynik testu językowego). Dodatkowo zastosowanie regresji logistycznej na walidacyjnym zbiorze danych pokazuje, że nasz model jest dokładny w prawie 80% co daje całkiem dobry wynik.

7. Źródła

- Skrypt do wykładu „Analiza danych jakościowych”, prof. dr hab. Zbigniew Szkutnik.
- Praca Licencjacka zamieszczona na stronie: http://www.biecek.pl/PASIK/uploads/JoannaGiemzaKatarzynaZwierzchowska.pdf