Data Science Language Analysis

Statistical Analysis

Nazli Ozum Kafaee
Prash Medirattaa
Avinash Prabhakaran

2018-04-15

```
#Loading the required packages
suppressPackageStartupMessages(library(tidyverse))
suppressPackageStartupMessages(library(knitr))
```

```
#Reading in processed data.
responses <- read.csv(file = "../docs/survey_results_clean.csv")
```

```
#preference encoding the response variable. Python -> 1; R -> 0
data <- responses %>% mutate(preference = if_else(preference == "Python", 1, 0))

data$background <- as.character(data$background)
data <- data %>%
  mutate(background = ifelse(background == "Computer Science / Computer Engineering",
                             "Computer Sc/Eng",
                              ifelse(background == "Mathematics / Statistics",
                                     "Maths/Stats", background)))

#Releveling the reference task from Data Viz -> Machine Learning
data_relevel <- data
data_relevel$task <-relevel(data$task,ref="Machine Learning")

#Fitting a GLM without any confounding variables.
mod <- glm(preference ~ task, family = binomial(link = 'logit'), data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = preference ~ task, family = binomial(link = "logit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9623  -0.8519   0.5615   0.5615   1.5425
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.8267     0.4532  -1.824   0.0681 .
## taskData wrangling     0.8267     0.7008   1.180   0.2381
## taskMachine Learning   2.5943     0.6104   4.250 2.14e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 110.372  on 84  degrees of freedom
## Residual deviance:  87.555  on 82  degrees of freedom
## AIC: 93.555
##
## Number of Fisher Scoring iterations: 4
```

```
#Fitting GLM with all the confounding variables.
mod <- glm(preference ~ task + background + experience + attitude + first + active,
           family = binomial(link = 'logit'), data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = preference ~ task + background + experience + attitude +
##     first + active, family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.90837  -0.37589   0.03797   0.36210   2.55585
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -5.3267     3.7988  -1.402   0.1609
## taskData wrangling            2.1788     1.2077   1.804   0.0712 .
## taskMachine Learning          3.5525     1.2092   2.938   0.0033 **
## backgroundComputer Sc/Eng     3.6830     1.7474   2.108   0.0351 *
## backgroundEngineering        -0.3603     1.5457  -0.233   0.8157
## backgroundMaths/Stats         0.6773     1.2335   0.549   0.5829
## backgroundOther               0.5810     1.4710   0.395   0.6929
## experienceLess than 1        -0.5551     0.9276  -0.598   0.5495
## experienceMore than 5        -3.0570     2.1747  -1.406   0.1598
## attitudeNo                   -0.9623     2.1569  -0.446   0.6555
## attitudeYes                   1.4492     1.2635   1.147   0.2514
## firstJava                    -2.0377     1.6539  -1.232   0.2179
## firstMatlab                   0.3319     1.6907   0.196   0.8444
## firstOther                    0.8743     1.3958   0.626   0.5311
## firstPython                   4.0073     2.0407   1.964   0.0496 *
## firstR                       -0.9978     1.5098  -0.661   0.5087
## firstSAS                     -0.7765     1.9379  -0.401   0.6886
## active2                       1.0669     2.7495   0.388   0.6980
## active3                       3.8766     3.0529   1.270   0.2042
## active4                       1.7934     3.1708   0.566   0.5717
## active5 or more              19.6718  2192.1364   0.009   0.9928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.372  on 84  degrees of freedom
## Residual deviance:  48.394  on 64  degrees of freedom
## AIC: 90.394
##
## Number of Fisher Scoring iterations: 17
```

```r
#Removing Attitude as Confounder as change
mod <- glm(preference ~ task + background + experience + first + active,
           family = binomial(link = 'logit'), data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = preference ~ task + background + experience + first +
##     active, family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6732  -0.3554   0.0497   0.4396   2.6319
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -4.5485     3.1287  -1.454 0.146006
## taskData wrangling            2.2772     1.1859   1.920 0.054824 .
## taskMachine Learning          4.1062     1.1574   3.548 0.000389 ***
## backgroundComputer Sc/Eng     3.9260     1.7344   2.264 0.023597 *
## backgroundEngineering        -0.7602     1.4649  -0.519 0.603812
## backgroundMaths/Stats         0.2319     1.1111   0.209 0.834691
## backgroundOther               0.7808     1.3918   0.561 0.574814
## experienceLess than 1        -0.7904     0.9347  -0.846 0.397758
## experienceMore than 5        -2.7743     2.0524  -1.352 0.176476
## firstJava                    -2.3143     1.5965  -1.450 0.147179
## firstMatlab                  -0.3195     1.6155  -0.198 0.843204
## firstOther                    0.9801     1.3456   0.728 0.466374
## firstPython                   3.2122     1.6282   1.973 0.048510 *
## firstR                       -1.0210     1.4712  -0.694 0.487702
## firstSAS                     -0.3420     1.8965  -0.180 0.856889
## active2                       1.3570     2.1491   0.631 0.527737
## active3                       4.0532     2.4874   1.629 0.103209
## active4                       2.4884     2.7204   0.915 0.360346
## active5 or more              19.6832  2173.2657   0.009 0.992774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.37  on 84  degrees of freedom
## Residual deviance:  50.42  on 66  degrees of freedom
## AIC: 88.42
##
## Number of Fisher Scoring iterations: 17
```

```r
#Removing Experience as Confounder
mod <- glm(preference ~ task + background + first + active,
           family = binomial(link = 'logit'), data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = preference ~ task + background + first + active,
##     family = binomial(link = "logit"), data = data)
```

```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.71388  -0.41985   0.05763   0.36812   2.86063
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -4.9757     2.9379  -1.694 0.090333 .
## taskData wrangling            2.1236     1.1197   1.897 0.057872 .
## taskMachine Learning          4.1142     1.1776   3.494 0.000476 ***
## backgroundComputer Sc/Eng     3.8721     1.6800   2.305 0.021178 *
## backgroundEngineering        -0.3729     1.2800  -0.291 0.770819
## backgroundMaths/Stats         0.4743     1.0138   0.468 0.639855
## backgroundOther               0.4334     1.2456   0.348 0.727903
## firstJava                    -2.0701     1.4490  -1.429 0.153113
## firstMatlab                  -0.6067     1.5278  -0.397 0.691302
## firstOther                    0.6676     1.2053   0.554 0.579665
## firstPython                   3.0143     1.6049   1.878 0.060356 .
## firstR                       -0.9333     1.4231  -0.656 0.511941
## firstSAS                     -0.6028     1.7588  -0.343 0.731806
## active2                       1.4009     2.1620   0.648 0.516992
## active3                       3.8922     2.3830   1.633 0.102411
## active4                       2.7216     2.6835   1.014 0.310493
## active5 or more              20.2089  2047.0820   0.010 0.992123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.372  on 84  degrees of freedom
## Residual deviance:  52.694  on 68  degrees of freedom
## AIC: 86.694
##
## Number of Fisher Scoring iterations: 17
```

```r
#Removing active as Confounder
mod <- glm(preference ~ task + background + first,
           family = binomial(link = 'logit'), data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = preference ~ task + background + first, family = binomial(link = "logit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1334  -0.5820   0.1590   0.5909   2.5368
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.97466    1.45591  -1.356 0.175003
## taskData wrangling          1.53852    0.96548   1.594 0.111042
## taskMachine Learning        3.23967    0.88929   3.643 0.000269 ***
## backgroundComputer Sc/Eng   3.09956    1.45163   2.135 0.032742 *
```

```
## backgroundEngineering      0.01428    1.20813   0.012 0.990568
## backgroundMaths/Stats      0.32001    0.96980   0.330 0.741417
## backgroundOther            0.58445    1.09830   0.532 0.594631
## firstJava                 -1.73120    1.25914  -1.375 0.169161
## firstMatlab               -0.29960    1.23445  -0.243 0.808236
## firstOther                 0.90223    1.05911   0.852 0.394286
## firstPython                1.50843    1.21850   1.238 0.215737
## firstR                    -1.50020    1.30055  -1.154 0.248703
## firstSAS                  -1.20225    1.66039  -0.724 0.469018
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.372  on 84  degrees of freedom
## Residual deviance:  64.428  on 72  degrees of freedom
## AIC: 90.428
##
## Number of Fisher Scoring iterations: 6
```

Not Removing active as the AIC score of the model increases from 86 to 90.

```
#Removing first as Confounder
mod <- glm(preference ~ task + background + active,
           family = binomial(link = 'logit'), data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = preference ~ task + background + active, family = binomial(link = "logit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8748  -0.4723   0.2731   0.7162   2.1546
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -1.5684     1.8226  -0.861   0.3895
## taskData wrangling           1.9769     0.9406   2.102   0.0356 *
## taskMachine Learning         3.3671     0.8528   3.948 7.88e-05 ***
## backgroundComputer Sc/Eng    2.6873     1.1777   2.282   0.0225 *
## backgroundEngineering       -0.2201     0.9315  -0.236   0.8132
## backgroundMaths/Stats        0.6468     0.8725   0.741   0.4585
## backgroundOther              0.5662     0.9426   0.601   0.5481
## active2                     -1.2156     1.7419  -0.698   0.4853
## active3                      0.4494     1.7776   0.253   0.8004
## active4                     -0.4845     2.0996  -0.231   0.8175
## active5 or more             17.4542  2141.6796   0.008   0.9935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.372  on 84  degrees of freedom
```

```
## Residual deviance:  68.411  on 74  degrees of freedom
## AIC: 90.411
##
## Number of Fisher Scoring iterations: 17
```

Not Removing `first language` as the AIC score of the model increases from 86 to 90.

```
#Removing background as Confounder
mod <- glm(preference ~ task + first + active,
           family = binomial(link = 'logit'), data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = preference ~ task + first + active, family = binomial(link = "logit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9176  -0.5017   0.2333   0.6315   2.6352
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.3523     1.6305  -0.829 0.406906
## taskData wrangling      1.1946     0.9287   1.286 0.198362
## taskMachine Learning    3.1200     0.8826   3.535 0.000408 ***
## firstJava              -1.5078     1.1569  -1.303 0.192484
## firstMatlab            -0.8223     1.3228  -0.622 0.534186
## firstOther             -0.2259     1.0561  -0.214 0.830614
## firstPython             1.7076     1.1688   1.461 0.144007
## firstR                 -1.8317     1.1531  -1.588 0.112182
## firstSAS               -1.5946     1.4252  -1.119 0.263195
## active2                -0.2567     1.4501  -0.177 0.859496
## active3                 1.8231     1.6670   1.094 0.274128
## active4                 1.3506     1.9144   0.706 0.480483
## active5 or more        18.2516  2317.0584   0.008 0.993715
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.372  on 84  degrees of freedom
## Residual deviance:  63.657  on 72  degrees of freedom
## AIC: 89.657
##
## Number of Fisher Scoring iterations: 17
```

Not Removing `background` as the AIC score of the model increases from 86 to 89.

```
#Model with first language, background and active
mod <- glm(preference ~ task + background + first + active,
           family = binomial(link = 'logit'), data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = preference ~ task + background + first + active,
```

```
##     family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.71388  -0.41985   0.05763   0.36812   2.86063
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -4.9757     2.9379  -1.694 0.090333 .
## taskData wrangling            2.1236     1.1197   1.897 0.057872 .
## taskMachine Learning          4.1142     1.1776   3.494 0.000476 ***
## backgroundComputer Sc/Eng     3.8721     1.6800   2.305 0.021178 *
## backgroundEngineering        -0.3729     1.2800  -0.291 0.770819
## backgroundMaths/Stats         0.4743     1.0138   0.468 0.639855
## backgroundOther               0.4334     1.2456   0.348 0.727903
## firstJava                    -2.0701     1.4490  -1.429 0.153113
## firstMatlab                  -0.6067     1.5278  -0.397 0.691302
## firstOther                    0.6676     1.2053   0.554 0.579665
## firstPython                   3.0143     1.6049   1.878 0.060356 .
## firstR                       -0.9333     1.4231  -0.656 0.511941
## firstSAS                     -0.6028     1.7588  -0.343 0.731806
## active2                       1.4009     2.1620   0.648 0.516992
## active3                       3.8922     2.3830   1.633 0.102411
## active4                       2.7216     2.6835   1.014 0.310493
## active5 or more              20.2089  2047.0820   0.010 0.992123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.372  on 84  degrees of freedom
## Residual deviance:  52.694  on 68  degrees of freedom
## AIC: 86.694
##
## Number of Fisher Scoring iterations: 17
```

```
#Releveled model with first language, background and active
model <- glm(preference ~ task + background + first + active,
             family = binomial(link = 'logit'), data = data_relevel)
summary(model)
```

```
##
## Call:
## glm(formula = preference ~ task + background + first + active,
##     family = binomial(link = "logit"), data = data_relevel)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.71388  -0.41985   0.05763   0.36812   2.86063
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -0.8615     2.6528  -0.325 0.745369
## taskData visualization       -4.1142     1.1776  -3.494 0.000476 ***
## taskData wrangling           -1.9906     0.9898  -2.011 0.044324 *
```

```
## backgroundComputer Sc/Eng    3.8721     1.6800   2.305 0.021178 *
## backgroundEngineering       -0.3729     1.2800  -0.291 0.770819
## backgroundMaths/Stats        0.4743     1.0138   0.468 0.639855
## backgroundOther              0.4334     1.2456   0.348 0.727903
## firstJava                   -2.0701     1.4490  -1.429 0.153113
## firstMatlab                 -0.6067     1.5278  -0.397 0.691302
## firstOther                   0.6676     1.2053   0.554 0.579665
## firstPython                  3.0143     1.6049   1.878 0.060356 .
## firstR                      -0.9333     1.4231  -0.656 0.511941
## firstSAS                    -0.6028     1.7588  -0.343 0.731806
## active2                      1.4009     2.1620   0.648 0.516992
## active3                      3.8922     2.3830   1.633 0.102411
## active4                      2.7216     2.6835   1.014 0.310493
## active5 or more             20.2089  2047.0820   0.010 0.992123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.372  on 84  degrees of freedom
## Residual deviance:  52.694  on 68  degrees of freedom
## AIC: 86.694
##
## Number of Fisher Scoring iterations: 17
```

```r
# Comparison of AIC scores for all the models
bind_cols(model = c("preference ~ task",
                    "preference ~ task + background + experience + attitude + first + active",
                    "preference ~ task + background + experience + first + active",
                    "preference ~ task + background + first + active",
                    "preference ~ task + background + first",
                    "preference ~ task + background + active",
                    "preference ~ task + first + active"),
          AIC = c(93.555, 90.394, 88.42, 86.694, 90.428, 90.411, 89.657)) %>%
  kable()
```

model	AIC
preference ~ task	93.555
preference ~ task + background + experience + attitude + first + active	90.394
preference ~ task + background + experience + first + active	88.420
preference ~ task + background + first + active	86.694
preference ~ task + background + first	90.428
preference ~ task + background + active	90.411
preference ~ task + first + active	89.657