Prediction Model

Aniol Garriga Torra

2023-11-12

To solve this problem we can consider that a good way to determine which factors determine whether students use public or private transportation systems to go to the university it will be to make a prediction model for the different factors of the student (the different questions of the survey) and look the explicative variables that have low p-value (this explains how important is this variable for the explain of our response variable).

```
df = read.csv('new_dataset.csv')
```

```
model1 <- glm(public_transport ~ sex + faculty + days + fastest + cheapest + most_comfortable + only_op
summary(model1)
```

```
##
## Call:
## glm(formula = public_transport ~ sex + faculty + days + fastest +
##     cheapest + most_comfortable + only_option + environment +
##     healthiest + no_private_vehicle, family = binomial, data = df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.675  -1.164   0.815   1.074   1.929
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.01468    0.29419  -0.050 0.960209
## sex                -0.22155    0.11350  -1.952 0.050929 .
## faculty            -0.02478    0.01133  -2.186 0.028794 *
## days                0.10114    0.04711   2.147 0.031806 *
## fastest            -0.09979    0.14384  -0.694 0.487849
## cheapest            0.42555    0.14392   2.957 0.003107 **
## most_comfortable   -0.25431    0.14397  -1.766 0.077324 .
## only_option         0.46415    0.19039   2.438 0.014772 *
## environment         0.22927    0.21391   1.072 0.283799
## healthiest         -1.17973    0.30488  -3.869 0.000109 ***
## no_private_vehicle -1.39952    0.28525  -4.906 9.28e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2004.5  on 1447  degrees of freedom
## Residual deviance: 1899.5  on 1437  degrees of freedom
## AIC: 1921.5
```

```
##
## Number of Fisher Scoring iterations: 4
```

In our first model that we take all the variables, we can see that the lowest important factor is the people their transport is the fastest option. We eliminate the variable.

```
model2 <- glm(public_transport ~ sex + faculty + days + cheapest + most_comfortable + only_option + envi
summary(model2)
```

```
##
## Call:
## glm(formula = public_transport ~ sex + faculty + days + cheapest +
##     most_comfortable + only_option + environment + healthiest +
##     no_private_vehicle, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6693  -1.1509   0.8197   1.0768   1.9398
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.11759    0.25398  -0.463 0.643374
## sex                -0.22497    0.11337  -1.984 0.047209 *
## faculty            -0.02462    0.01133  -2.173 0.029756 *
## days                0.10061    0.04710   2.136 0.032669 *
## cheapest            0.46385    0.13296   3.489 0.000486 ***
## most_comfortable   -0.22516    0.13761  -1.636 0.101800
## only_option         0.53251    0.16303   3.266 0.001089 **
## environment         0.28315    0.19924   1.421 0.155279
## healthiest         -1.12167    0.29306  -3.827 0.000129 ***
## no_private_vehicle -1.35080    0.27628  -4.889 1.01e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2004.5  on 1447  degrees of freedom
## Residual deviance: 1900.0  on 1438  degrees of freedom
## AIC: 1920
##
## Number of Fisher Scoring iterations: 4
```

In this second model we can see that the lowest important variable is the people who only cares the environment. Counterintuitive no?

```
model3 <- glm(public_transport ~ sex + faculty + days + cheapest + most_comfortable + only_option + heal
summary(model3)
```

```
##
## Call:
## glm(formula = public_transport ~ sex + faculty + days + cheapest +
##     most_comfortable + only_option + healthiest + no_private_vehicle,
##     family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6399  -1.1569   0.8296   1.0767   1.9487
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.07384    0.25195  -0.293 0.769456
## sex                -0.22336    0.11328  -1.972 0.048642 *
## faculty            -0.02500    0.01132  -2.208 0.027244 *
## days                0.09996    0.04705   2.124 0.033632 *
## cheapest            0.46387    0.13284   3.492 0.000479 ***
## most_comfortable   -0.25308    0.13619  -1.858 0.063132 .
## only_option         0.50131    0.16156   3.103 0.001916 **
## healthiest         -1.09099    0.29202  -3.736 0.000187 ***
## no_private_vehicle -1.38441    0.27530  -5.029 4.94e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2004.5  on 1447  degrees of freedom
## Residual deviance: 1902.0  on 1439  degrees of freedom
## AIC: 1920
##
## Number of Fisher Scoring iterations: 4
```

We eliminate the variable that explains the people who only cares if is the most confortable way to travel.

```
model4 <- glm(public_transport ~ sex + faculty + days + cheapest + only_option + healthiest + no_privat
summary(model4)
```

```
##
## Call:
## glm(formula = public_transport ~ sex + faculty + days + cheapest +
##     only_option + healthiest + no_private_vehicle, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6768  -1.1722   0.8251   1.0907   1.8711
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.21958    0.23914  -0.918 0.358504
## sex                -0.22403    0.11314  -1.980 0.047686 *
## faculty            -0.02492    0.01131  -2.203 0.027601 *
## days                0.10032    0.04695   2.137 0.032604 *
## cheapest            0.56847    0.12051   4.717 2.39e-06 ***
## only_option         0.62266    0.14797   4.208 2.57e-05 ***
## healthiest         -0.99931    0.28794  -3.471 0.000519 ***
## no_private_vehicle -1.29322    0.27091  -4.774 1.81e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2004.5  on 1447  degrees of freedom
## Residual deviance: 1905.5  on 1440  degrees of freedom
## AIC: 1921.5
##
## Number of Fisher Scoring iterations: 4
```

The sex of the person doesn’t matter…

```
model5 <- glm(public_transport ~ faculty + days + cheapest + only_option + healthiest + no_private_vehi
summary(model5)
```

```
##
## Call:
## glm(formula = public_transport ~ faculty + days + cheapest +
##     only_option + healthiest + no_private_vehicle, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7176  -1.1833   0.8521   1.1126   1.9251
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.36752    0.22712  -1.618 0.105617
## faculty            -0.02366    0.01128  -2.098 0.035871 *
## days                0.09992    0.04691   2.130 0.033155 *
## cheapest            0.56452    0.12033   4.692 2.71e-06 ***
## only_option         0.63709    0.14774   4.312 1.62e-05 ***
## healthiest         -0.99703    0.28740  -3.469 0.000522 ***
## no_private_vehicle -1.29663    0.27059  -4.792 1.65e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2004.5  on 1447  degrees of freedom
## Residual deviance: 1909.4  on 1441  degrees of freedom
## AIC: 1923.4
##
## Number of Fisher Scoring iterations: 4
```

The faculty that the student studies doesn’t really matter.

```
model6 <- glm(public_transport ~ days + cheapest + only_option + healthiest + no_private_vehicle, data
summary(model6)
```

```
##
```

```
## Call:
## glm(formula = public_transport ~ days + cheapest + only_option +
##     healthiest + no_private_vehicle, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6974  -1.1750   0.8737   1.1199   1.9309
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.52808    0.21377  -2.470 0.013500 *
## days                0.10448    0.04679   2.233 0.025538 *
## cheapest            0.56046    0.12012   4.666 3.07e-06 ***
## only_option         0.61563    0.14709   4.185 2.85e-05 ***
## healthiest         -0.99217    0.28695  -3.458 0.000545 ***
## no_private_vehicle -1.27225    0.27012  -4.710 2.48e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2004.5  on 1447  degrees of freedom
## Residual deviance: 1913.8  on 1442  degrees of freedom
## AIC: 1925.8
##
## Number of Fisher Scoring iterations: 4
```

And finally we can say that the days that students goes to university doesn’t matter.

```
model7 <- glm(public_transport ~ cheapest + only_option + healthiest + no_private_vehicle, data = df, fa
summary(model7)
```

```
##
## Call:
## glm(formula = public_transport ~ cheapest + only_option + healthiest +
##     no_private_vehicle, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6616  -1.1424   0.9571   0.9874   1.7972
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.08285    0.07625  -1.087 0.277224
## cheapest            0.54762    0.11976   4.573 4.81e-06 ***
## only_option         0.62603    0.14683   4.264 2.01e-05 ***
## healthiest         -0.98866    0.28659  -3.450 0.000561 ***
## no_private_vehicle -1.31045    0.26912  -4.869 1.12e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 2004.5  on 1447  degrees of freedom
## Residual deviance: 1918.8  on 1443  degrees of freedom
## AIC: 1928.8
##
## Number of Fisher Scoring iterations: 4
```

Finally we can see that the group of this factors explains pretty good our variable response. The conclusion it will be that:

By the methodology of the statistic binomial model and with the data of the UPC student's survey, we can say that the most important factors for determine whether students use public or private transportation systems to go to the university, are:

-If the student take the cheapest option

-If the student take the only option he has

-If the student take the healthiest option

-If the student doesn't have private vehicle