**#Importing dataset**

> library(readr)

> ridesharedata <- read\_csv("DATA/ridesharedata (1).csv")

Rows: 15892 Columns: 7

> View(ridesharedata)

> ridesharedata<-read.csv(file.choose())

**#exploring the dataset**

> str(ridesharedata)

'data.frame': 15892 obs. of 7 variables:

$ workingdaysperweek : int 5 5 5 5 5 5 5 5 5 4 ...

$ distanceinkmsforcommute: num 18.6 16.1 17.1 21.3 20.3 ...

$ usedrideshare : chr "No" "No" "No" "No" ...

$ gender : chr "Female" "Female" "Female" "Female" ...

$ salary : int 8496 6521 3643 7059 6055 3954 3296 7543 3339 387 ...

$ itwork : int 2 2 1 2 1 2 1 2 2 1 ...

$ workexp : int 4 1 3 2 5 2 2 4 2 5 ...

**#Importing packages**

> install.packages("pacman")

Error in install.packages : Updating loaded packages

> pacman::p\_load(ggplot2, dplyr, reader, caret, psych, MASS, descr, boot)

> install.packages("pacman")

**# count of the dependent variable**

> summary(ridesharedata)

workingdaysperweek distanceinkmsforcommute usedrideshare gender salary itwork

Min. :1.000 Min. :10.00 Length:15892 Length:15892 Min. : 18 Min. :1.000

1st Qu.:5.000 1st Qu.:15.52 Class :character Class :character 1st Qu.: 4187 1st Qu.:1.000

Median :5.000 Median :21.06 Mode :character Mode :character Median : 6778 Median :2.000

Mean :4.874 Mean :21.40 Mean : 6875 Mean :1.503

3rd Qu.:5.000 3rd Qu.:26.53 3rd Qu.: 9154 3rd Qu.:2.000

Max. :7.000 Max. :39.99 Max. :15334 Max. :2.000

workexp

Min. :1.000

1st Qu.:2.000

Median :3.000

Mean :3.502

3rd Qu.:5.000

Max. :6.000

> str(ridesharedata, give.attr = FALSE)

'data.frame': 15892 obs. of 7 variables:

$ workingdaysperweek : int 5 5 5 5 5 5 5 5 5 4 ...

$ distanceinkmsforcommute: num 18.6 16.1 17.1 21.3 20.3 ...

$ usedrideshare : chr "No" "No" "No" "No" ...

$ gender : chr "Female" "Female" "Female" "Female" ...

$ salary : int 8496 6521 3643 7059 6055 3954 3296 7543 3339 387 ...

$ itwork : int 2 2 1 2 1 2 1 2 2 1 ...

$ workexp : int 4 1 3 2 5 2 2 4 2 5 ...

> ggplot(ridesharedata, aes(x = usedrideshare)) + geom\_histogram(stat = "count")

Warning message:

In geom\_histogram(stat = "count") :

Ignoring unknown parameters: `binwidth`, `bins`, and `pad`

**#encoding the dependent variable to factor**

> ridesharedata$usedrideshare <- factor(ridesharedata$usedrideshare,levels = c("No","Yes"),

+ labels = c(0,1))

**#changing the categorical labels to factors**

> ridesharedata$workingdaysperweek <- as.factor(ridesharedata$workingdaysperweek)

> ridesharedata$gender <- factor(ridesharedata$gender,levels = c("Female","Male"),

+ labels = c(0,1))

> ridesharedata$itwork <- as.factor(ridesharedata$itwork)

> ridesharedata$workexp <- as.factor(ridesharedata$workexp)

**#Building a full logistic regression model with all available predictors**

> logitModelFull <- glm(usedrideshare ~ workingdaysperweek+distanceinkmsforcommute+gender+salary+itwork+workexp, family = binomial,

+ data = ridesharedata)

> LogRegR2(logitModelFull)

Chi2 8589.816

Df 15

Sig. 0

Cox and Snell Index 0.41755

Nagelkerke Index 0.589683

McFadden's R2 0.4389703

> summary(logitModelFull)

Call:

glm(formula = usedrideshare ~ workingdaysperweek + distanceinkmsforcommute +

gender + salary + itwork + workexp, family = binomial, data = ridesharedata)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.9169 -0.5677 -0.2085 0.3013 3.4445

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.29189244 0.35961293 -23.058 < 0.0000000000000002 \*\*\*

workingdaysperweek2 0.18731646 0.40963049 0.457 0.647469

workingdaysperweek3 0.12571477 0.35111887 0.358 0.720313

workingdaysperweek4 0.50843958 0.33939945 1.498 0.134119

workingdaysperweek5 0.50615679 0.33014666 1.533 0.125245

workingdaysperweek6 0.68218589 0.34504211 1.977 0.048029 \*

workingdaysperweek7 0.85961028 0.35931783 2.392 0.016741 \*

distanceinkmsforcommute 0.09733786 0.00356006 27.342 < 0.0000000000000002 \*\*\*

gender1 -0.18436666 0.05049240 -3.651 0.000261 \*\*\*

salary 0.00063337 0.00001112 56.947 < 0.0000000000000002 \*\*\*

itwork2 -0.04531049 0.04804190 -0.943 0.345607

workexp2 -0.00996592 0.08373250 -0.119 0.905259

workexp3 0.00981934 0.08225385 0.119 0.904975

workexp4 -0.10176995 0.08293352 -1.227 0.219775

workexp5 -0.07432350 0.08356044 -0.889 0.373757

workexp6 -0.04836414 0.08308208 -0.582 0.560483

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 19568 on 15891 degrees of freedom

Residual deviance: 10978 on 15876 degrees of freedom

AIC: 11010

Number of Fisher Scoring iterations: 6

**# INTERPRETATION OF THE ABOVE RESULTS**

Two things we are interested from the above results are the ‘sign’ of the coefficients and the significance values. The values of the coefficients are not useful in from the results of a Logistic Regression model. Based on that, if we consider anything with 0.05 significance value as significant;

a)wokingdaysperweek6 & 7 are positively affecting the likelihood of using the ride share. Which means that people working 6 or 7 days per week or more likely to use the ride share.

b)’distanceinkmsforcommute’ is also positively affecting the dependent variable. More the distance in commute, more likely to use rideshare.

c)’gender1’ (Male) is negatively affecting the dependent variable. Meaning gender ‘Female’ is more likely to use the rideshare than males.

d)Salary is positively affecting the outcome. Meaning, higher the Salary the more likelihood of using rideshare.

Rest of the variables are not significant enough.

**#checking the pseudo R-Squared**

> LogRegR2(logitModelFull)

Chi2 8589.816

Df 15

Sig. 0

Cox and Snell Index 0.41755

Nagelkerke Index 0.589683

McFadden's R2 0.4389703

> summary(logitModelFull)

Call:

glm(formula = usedrideshare ~ workingdaysperweek + distanceinkmsforcommute +

gender + salary + itwork + workexp, family = binomial, data = ridesharedata)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.9169 -0.5677 -0.2085 0.3013 3.4445

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.29189244 0.35961293 -23.058 < 0.0000000000000002 \*\*\*

workingdaysperweek2 0.18731646 0.40963049 0.457 0.647469

workingdaysperweek3 0.12571477 0.35111887 0.358 0.720313

workingdaysperweek4 0.50843958 0.33939945 1.498 0.134119

workingdaysperweek5 0.50615679 0.33014666 1.533 0.125245

workingdaysperweek6 0.68218589 0.34504211 1.977 0.048029 \*

workingdaysperweek7 0.85961028 0.35931783 2.392 0.016741 \*

distanceinkmsforcommute 0.09733786 0.00356006 27.342 < 0.0000000000000002 \*\*\*

gender1 -0.18436666 0.05049240 -3.651 0.000261 \*\*\*

salary 0.00063337 0.00001112 56.947 < 0.0000000000000002 \*\*\*

itwork2 -0.04531049 0.04804190 -0.943 0.345607

workexp2 -0.00996592 0.08373250 -0.119 0.905259

workexp3 0.00981934 0.08225385 0.119 0.904975

workexp4 -0.10176995 0.08293352 -1.227 0.219775

workexp5 -0.07432350 0.08356044 -0.889 0.373757

workexp6 -0.04836414 0.08308208 -0.582 0.560483

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 19568 on 15891 degrees of freedom

Residual deviance: 10978 on 15876 degrees of freedom

AIC: 11010

Number of Fisher Scoring iterations: 6

> #verifying the model is still available

> summary(logitModelFull)

Call:

glm(formula = usedrideshare ~ workingdaysperweek + distanceinkmsforcommute +

gender + salary + itwork + workexp, family = binomial, data = ridesharedata)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.9169 -0.5677 -0.2085 0.3013 3.4445

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.29189244 0.35961293 -23.058 < 0.0000000000000002 \*\*\*

workingdaysperweek2 0.18731646 0.40963049 0.457 0.647469

workingdaysperweek3 0.12571477 0.35111887 0.358 0.720313

workingdaysperweek4 0.50843958 0.33939945 1.498 0.134119

workingdaysperweek5 0.50615679 0.33014666 1.533 0.125245

workingdaysperweek6 0.68218589 0.34504211 1.977 0.048029 \*

workingdaysperweek7 0.85961028 0.35931783 2.392 0.016741 \*

distanceinkmsforcommute 0.09733786 0.00356006 27.342 < 0.0000000000000002 \*\*\*

gender1 -0.18436666 0.05049240 -3.651 0.000261 \*\*\*

salary 0.00063337 0.00001112 56.947 < 0.0000000000000002 \*\*\*

itwork2 -0.04531049 0.04804190 -0.943 0.345607

workexp2 -0.00996592 0.08373250 -0.119 0.905259

workexp3 0.00981934 0.08225385 0.119 0.904975

workexp4 -0.10176995 0.08293352 -1.227 0.219775

workexp5 -0.07432350 0.08356044 -0.889 0.373757

workexp6 -0.04836414 0.08308208 -0.582 0.560483

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 19568 on 15891 degrees of freedom

Residual deviance: 10978 on 15876 degrees of freedom

AIC: 11010

Number of Fisher Scoring iterations: 6

> coefsexp <- coef(logitModelFull) %>% exp() %>% round(2)

> coefsexp

(Intercept) workingdaysperweek2 workingdaysperweek3 workingdaysperweek4 workingdaysperweek5

0.00 1.21 1.13 1.66 1.66

workingdaysperweek6 workingdaysperweek7 distanceinkmsforcommute gender1 salary

1.98 2.36 1.10 0.83 1.00

itwork2 workexp2 workexp3 workexp4 workexp5

0.96 0.99 1.01 0.90 0.93

workexp6

0.95

# Interpretation of the odds ratio from above. The general rule is the odds ratio is that if it is more than 1, it is more likely to be a predictor. Based on that, ‘workingdaysperweek6 & 7’ have odds ratio of 1.98 and 2.36 respectively. Which means they are likely to predict if the person will use the rideshare. So is ‘distanceinkmsforcommute’ with 1.10 odds ratio. Therefore, a possible predictor. However, gender1 (male) is not a likely predictor. Of the rest, ‘salary’ & ‘workexp3’ cannot be taken as likely predictors of the dependent variable, as they have odds ratio of 1 & 1.01 respectively.

**#building a new aicmodel**

logitModelNew <- stepAIC(logitModelFull,trace = 0)

> summary(logitModelNew)

Call:

glm(formula = usedrideshare ~ workingdaysperweek + distanceinkmsforcommute +

gender + salary, family = binomial, data = ridesharedata)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.9422 -0.5686 -0.2085 0.3015 3.4452

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.34348840 0.35515422 -23.493 < 0.0000000000000002 \*\*\*

workingdaysperweek2 0.18471011 0.40984653 0.451 0.652219

workingdaysperweek3 0.11899265 0.35159985 0.338 0.735038

workingdaysperweek4 0.50328318 0.33989063 1.481 0.138681

workingdaysperweek5 0.50105514 0.33067293 1.515 0.129707

workingdaysperweek6 0.67464020 0.34552688 1.952 0.050879 .

workingdaysperweek7 0.85301585 0.35978254 2.371 0.017744 \*

distanceinkmsforcommute 0.09727394 0.00355817 27.338 < 0.0000000000000002 \*\*\*

gender1 -0.18425809 0.05047392 -3.651 0.000262 \*\*\*

salary 0.00063320 0.00001112 56.948 < 0.0000000000000002 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 19568 on 15891 degrees of freedom

Residual deviance: 10982 on 15882 degrees of freedom

AIC: 11002

Number of Fisher Scoring iterations: 6

**#saving the formula**

> formulaLogit <- as.formula(summary(logitModelNew)$call)

> formulaLogit

usedrideshare ~ workingdaysperweek + distanceinkmsforcommute +

gender + salary

As we can see only the significant variables are showing up.

# To find the odds ratio to check for changes  
coefsexp2 <- coef(logitModelNew) %>% exp() %>% round(2)  
coefsexp2

(Intercept) workingdaysperweek2 workingdaysperweek3 workingdaysperweek4 workingdaysperweek5

0.00 1.20 1.13 1.65 1.65

workingdaysperweek6 workingdaysperweek7 distanceinkmsforcommute gender1 salary

1.96 2.35 1.10 0.83 1.00

>

Based on the odds ratio after tuning the model we can conclude ‘workingdaysperweek 6&7’, ‘distanceinkms’,gender (Female) can be taken as predictors as they have 1.96,2.35,1..10 & 0.83(for males) respectively.

Therefore, to conclude, the variables I would include to build a model would be employees working 6 or 7 days a week, female, ‘distanceinkmsfor commute’. Female employees are more likely to have used the rideshare service. One of the limitations of the model, is that odds ratio for many ‘significant’ variables are only just about likely to be predictors, in the sense, there are no clear winners.