a3_1896845

Vineeth Marikuntematha Ravisharadhya

2024-04-21

```
#Loading the required library packages

library(here)
library(readr)
library(tidyverse)
library(dplyr)
library(inspectdf)
library(stringr)
library(ggplot2)
library(caret)
library(e1071)
library(tidymodels)
library(forcats)
library(rsample)
library(modelr)
library(parsnip)
library(car)
library(yardstick)
```

```
# As per the deliverable specification point 1 setting ysn and x formula as provided.
# Student number goes here
ysn = 1896845
# Calculating student number plus 2 modulo 3

file_num_ = (ysn + 2) %% 3
file_num_
```

```
## [1] 1
```

```
file_name_ <- paste0("./data/survey_",file_num_,".csv")
file_name_
```

```
## [1] "./data/survey_1.csv"
```

1. Load the correct dataset and save it as a tibble. Output the first 10 lines of the dataset.

```r
# Read the data and converting it to tibble
surv_ = read_csv(here::here("./data/survey_1.csv"))
surv_ =as_tibble(surv_)
surv_
```

```
## # A tibble: 50,000 x 7
##    recommend   age company_aware malfunction multi_purch SES   social_media
##        <dbl> <dbl> <lgl>         <lgl>       <lgl>       <chr> <lgl>
##  1         0    51 TRUE          TRUE        TRUE        low   TRUE
##  2         1    51 FALSE         FALSE       TRUE        high  TRUE
##  3         0    32 TRUE          TRUE        TRUE        high  FALSE
##  4         1    48 TRUE          FALSE       TRUE        high  TRUE
##  5         0    28 TRUE          FALSE       FALSE       low   TRUE
##  6         0    53 FALSE         FALSE       FALSE       low   FALSE
##  7         1    32 FALSE         FALSE       FALSE       low   TRUE
##  8         0    39 TRUE          FALSE       FALSE       mid   TRUE
##  9         0    27 FALSE         FALSE       FALSE       low   TRUE
## 10         1    49 TRUE          FALSE       TRUE        high  TRUE
## # i 49,990 more rows
```

```r
#gadget_data
head(surv_,10)
```

```
## # A tibble: 10 x 7
##    recommend   age company_aware malfunction multi_purch SES   social_media
##        <dbl> <dbl> <lgl>         <lgl>       <lgl>       <chr> <lgl>
##  1         0    51 TRUE          TRUE        TRUE        low   TRUE
##  2         1    51 FALSE         FALSE       TRUE        high  TRUE
##  3         0    32 TRUE          TRUE        TRUE        high  FALSE
##  4         1    48 TRUE          FALSE       TRUE        high  TRUE
##  5         0    28 TRUE          FALSE       FALSE       low   TRUE
##  6         0    53 FALSE         FALSE       FALSE       low   FALSE
##  7         1    32 FALSE         FALSE       FALSE       low   TRUE
##  8         0    39 TRUE          FALSE       FALSE       mid   TRUE
##  9         0    27 FALSE         FALSE       FALSE       low   TRUE
## 10         1    49 TRUE          FALSE       TRUE        high  TRUE
```

Q2. Using dot points, identify what types of variables we now have in our data set, i.e., "Quantitative Discrete",

"Quantitative Continuous", "Categorical Nominal", "Categorical Ordinal". (Don't just describe what data

type they are in the tibble — you need to think about the type of variable in the context of the meaning of

the data.) Make sure you provide some justification for your choice of variable types.

- recommend: The type is Categorical Nominal This column is considered as a Categorical Nominal as it is only representing two categories 0 and 1 , also there is neither levels nor rankings for classification here.
- age: The type of data type of this column is Quantitative continuous As the column name or variable is filled with ages of people , having decimal values as well, which is continuous. Therefore the age variable is a quantitative continuous variable.
- company_aware: The type is Categorical nominal This column is considered as Categorical nominal as it is representing only two categories TRUE or FALSE, showing whether the people are aware of the companies existence.Also there is no levels of classification here ,therefore is a categorical nominal variable.
- malfunction: The type of this is Categorical nominal the column is considered as Categorical nominal as it just represent only two categories either TRUE or FALSE, showing whether one of their gadget has been malfunctioned or not. Since there is no ranking or levels of classification here, this column is considered as a categorical nominal variable.
- multi_purch: The type of data type is Categorical nominal as it also represents only two categories that is TRUE and FALSE, showing whether the people did multiple purchase or not. Also there is no ranking or levels of classification here.
- SES: The type is Categorical Ordinal is considered as a categorical ordinal column because it mentions ranking of the social status of people with high being the highest level and low being the lowest level.
- social_media: This is considered as Categorical nominal as it just represent only two categories that is TRUE or FALSE, showing whether people are active in social media or not.

```
colnames(surv_)
```

```
## [1] "recommend"     "age"           "company_aware" "malfunction"
## [5] "multi_purch"   "SES"           "social_media"
```

Q3. Now it's time to tame our data. But since we are going to fit a logistic regression model, we need to modify

our requirements a little bit.

- **(a) Make sure that all column names are in snake case.**

- **(b) Make the variables age, company aware, malfunction, multi purch and social media conform to**

the Tame Data conventions in Module 2 (page 3).

- **(c) Convert recommend to a data type, with yes for 1 and no for 0.**

- **(d) Convert the Socio-Economic Status to a .**

(e) Output the first 10 rows of your data.

```
# as the  ses is in block letters we will convert column name to snake case ,
surv_= surv_ %>% rename(
  ses = SES)
surv_
```

```
## # A tibble: 50,000 x 7
##    recommend   age company_aware malfunction multi_purch ses   social_media
##        <dbl> <dbl> <lgl>         <lgl>       <lgl>       <chr> <lgl>
##  1         0    51 TRUE          TRUE        TRUE        low   TRUE
##  2         1    51 FALSE         FALSE       TRUE        high  TRUE
##  3         0    32 TRUE          TRUE        TRUE        high  FALSE
##  4         1    48 TRUE          FALSE       TRUE        high  TRUE
##  5         0    28 TRUE          FALSE       FALSE       low   TRUE
##  6         0    53 FALSE         FALSE       FALSE       low   FALSE
##  7         1    32 FALSE         FALSE       FALSE       low   TRUE
##  8         0    39 TRUE          FALSE       FALSE       mid   TRUE
##  9         0    27 FALSE         FALSE       FALSE       low   TRUE
## 10         1    49 TRUE          FALSE       TRUE        high  TRUE
## # i 49,990 more rows
```

```
surv_=relocate (surv_,"age", .before=recommend)
surv_
```

```
## # A tibble: 50,000 x 7
##      age recommend company_aware malfunction multi_purch ses   social_media
```

```
##    <dbl>     <dbl> <lgl>         <lgl>       <lgl>       <chr> <lgl>
##  1    51         0 TRUE          TRUE        TRUE        low   TRUE
##  2    51         1 FALSE         FALSE       TRUE        high  TRUE
##  3    32         0 TRUE          TRUE        TRUE        high  FALSE
##  4    48         1 TRUE          FALSE       TRUE        high  TRUE
##  5    28         0 TRUE          FALSE       FALSE       low   TRUE
##  6    53         0 FALSE         FALSE       FALSE       low   FALSE
##  7    32         1 FALSE         FALSE       FALSE       low   TRUE
##  8    39         0 TRUE          FALSE       FALSE       mid   TRUE
##  9    27         0 FALSE         FALSE       FALSE       low   TRUE
## 10    49         1 TRUE          FALSE       TRUE        high  TRUE
## # i 49,990 more rows
```

```
surv_ = surv_ %>% mutate (
  company_aware=as_factor(company_aware),
  malfunction=as_factor(malfunction),
  multi_purch=as_factor(multi_purch),
  social_media=as_factor(social_media))
surv_
```

```
## # A tibble: 50,000 x 7
##      age recommend company_aware malfunction multi_purch ses   social_media
##    <dbl>     <dbl> <fct>         <fct>       <fct>       <chr> <fct>
##  1    51         0 TRUE          TRUE        TRUE        low   TRUE
##  2    51         1 FALSE         FALSE       TRUE        high  TRUE
##  3    32         0 TRUE          TRUE        TRUE        high  FALSE
##  4    48         1 TRUE          FALSE       TRUE        high  TRUE
##  5    28         0 TRUE          FALSE       FALSE       low   TRUE
##  6    53         0 FALSE         FALSE       FALSE       low   FALSE
##  7    32         1 FALSE         FALSE       FALSE       low   TRUE
##  8    39         0 TRUE          FALSE       FALSE       mid   TRUE
##  9    27         0 FALSE         FALSE       FALSE       low   TRUE
## 10    49         1 TRUE          FALSE       TRUE        high  TRUE
## # i 49,990 more rows
```

```
#
surv_$ses <-factor(surv_$ses, levels = c("high", "mid", "low"))
surv_
```

```
## # A tibble: 50,000 x 7
##      age recommend company_aware malfunction multi_purch ses   social_media
##    <dbl>     <dbl> <fct>         <fct>       <fct>       <fct> <fct>
##  1    51         0 TRUE          TRUE        TRUE        low   TRUE
##  2    51         1 FALSE         FALSE       TRUE        high  TRUE
##  3    32         0 TRUE          TRUE        TRUE        high  FALSE
##  4    48         1 TRUE          FALSE       TRUE        high  TRUE
##  5    28         0 TRUE          FALSE       FALSE       low   TRUE
##  6    53         0 FALSE         FALSE       FALSE       low   FALSE
##  7    32         1 FALSE         FALSE       FALSE       low   TRUE
##  8    39         0 TRUE          FALSE       FALSE       mid   TRUE
##  9    27         0 FALSE         FALSE       FALSE       low   TRUE
## 10    49         1 TRUE          FALSE       TRUE        high  TRUE
## # i 49,990 more rows
```

```r
#So as the question 3 mandates to tame the data ,as we did on tame for ses we will do it for recommend
surv_ = surv_ %>%  mutate (recommend=as_factor(recommend))
surv_$recommend = fct_recode(surv_$recommend, "yes"="1","no"="0")
surv_
```

```
## # A tibble: 50,000 x 7
##      age recommend company_aware malfunction multi_purch ses   social_media
##    <dbl> <fct>     <fct>         <fct>       <fct>       <fct> <fct>
##  1    51 no        TRUE          TRUE        TRUE        low   TRUE
##  2    51 yes       FALSE         FALSE       TRUE        high  TRUE
##  3    32 no        TRUE          TRUE        TRUE        high  FALSE
##  4    48 yes       TRUE          FALSE       TRUE        high  TRUE
##  5    28 no        TRUE          FALSE       FALSE       low   TRUE
##  6    53 no        FALSE         FALSE       FALSE       low   FALSE
##  7    32 yes       FALSE         FALSE       FALSE       low   TRUE
##  8    39 no        TRUE          FALSE       FALSE       mid   TRUE
##  9    27 no        FALSE         FALSE       FALSE       low   TRUE
## 10    49 yes       TRUE          FALSE       TRUE        high  TRUE
## # i 49,990 more rows
```

```r
# Display the first 10 lines of the data
head(surv_,10)
```

```
## # A tibble: 10 x 7
##      age recommend company_aware malfunction multi_purch ses   social_media
##    <dbl> <fct>     <fct>         <fct>       <fct>       <fct> <fct>
##  1    51 no        TRUE          TRUE        TRUE        low   TRUE
##  2    51 yes       FALSE         FALSE       TRUE        high  TRUE
##  3    32 no        TRUE          TRUE        TRUE        high  FALSE
##  4    48 yes       TRUE          FALSE       TRUE        high  TRUE
##  5    28 no        TRUE          FALSE       FALSE       low   TRUE
##  6    53 no        FALSE         FALSE       FALSE       low   FALSE
##  7    32 yes       FALSE         FALSE       FALSE       low   TRUE
##  8    39 no        TRUE          FALSE       FALSE       mid   TRUE
##  9    27 no        FALSE         FALSE       FALSE       low   TRUE
## 10    49 yes       TRUE          FALSE       TRUE        high  TRUE
```

Q4.Setting the correct seed, split your data into a training set (with 40,000 rows) and a testing set, with the

remaining rows. Use the command dim() to output the dimensions of your training and testing sets.

```r
#setting the seed as per the deliverable specification point 2.
set.seed(1896845)

#split the data to training and testing data set
surv_splt_ = initial_split( surv_,prop=0.8 )
```

```
surv_tn_ = training( surv_splt_ )
surv_tt_ = testing( surv_splt_ )

# Output the dimensions
dim(surv_tn_)
```

```
## [1] 40000     7
```

```
dim(surv_tt_)
```

```
## [1] 10000     7
```

Q5. Fit a logistic regression model to your training data, with recommend as the response and all other variables

as the predictors. Output the summary of the model.

```
#Using logistic regression model for training data.

surv_tn_logrr_ <- logistic_reg() %>% set_engine("glm")
surv_ft_ <- surv_tn_logrr_ %>%
fit(recommend ~ ., data = surv_tn_)

# Lets check for the summary of the model and get all the variables.
summary(surv_ft_$fit)
```

```
##
## Call:
## stats::glm(formula = recommend ~ ., family = stats::binomial,
##     data = data)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.006192   0.076093  -0.081    0.935
## age               -0.051199   0.001395 -36.697   <2e-16 ***
## company_awareTRUE -0.047023   0.030164  -1.559    0.119
## malfunctionTRUE   -5.888283   0.207368 -28.395   <2e-16 ***
## multi_purchTRUE    3.216367   0.033144  97.042   <2e-16 ***
## sesmid             0.030293   0.036724   0.825    0.409
## seslow             0.378270   0.036233  10.440   <2e-16 ***
## social_mediaTRUE  -0.040082   0.040268  -0.995    0.320
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47800  on 39999  degrees of freedom
## Residual deviance: 29426  on 39992  degrees of freedom
```

```
## AIC: 29442
##
## Number of Fisher Scoring iterations: 8
```

Q6. Use the command model matrix() on the ses variable of your training data to see what happens to ses

when we fit a model. (See pages 2 – 6 in Module 7.)

```
# creating matrix model for ses as per the module 7, page 3.
mdl_mat = model_matrix(surv_tn_,~ses)
mdl_mat
```

```
## # A tibble: 40,000 x 3
##    `(Intercept)` sesmid seslow
##            <dbl>  <dbl>  <dbl>
##  1             1      0      1
##  2             1      0      0
##  3             1      1      0
##  4             1      0      0
##  5             1      1      0
##  6             1      0      0
##  7             1      0      1
##  8             1      0      1
##  9             1      0      0
## 10             1      0      0
## # i 39,990 more rows
```

(a) How many new variables have been introduced?

- 2 new columns has been introduced. One column seshigh for whether the person has high social economic status. And another column seslow for whether the person has low social economic status or not.

(b) What is the reference level for ses?

- The value mid(middle or medium as per the data instruction in the handout of the client) is considered as the reference for ses.

Q7.(a) Build a new tibble called ses matrix, with the first column giving the true ses data, and the second and

third columns giving the coordinates of the ses value in the new variables defined for the ses variable.

Call these new variables seslow and sesmid. (It should be clear which one is which.)

```
ses_mx = tibble(
  ses = surv_tn_$ses,
  seslow = mdl_mat$seslow,
  sesmid = mdl_mat$sesmid
)
ses_mx
```

```
## # A tibble: 40,000 x 3
##    ses   seslow sesmid
##    <fct>  <dbl>  <dbl>
##  1 low        1      0
##  2 high       0      0
##  3 mid        0      1
##  4 high       0      0
##  5 mid        0      1
##  6 high       0      0
##  7 low        1      0
##  8 low        1      0
##  9 high       0      0
## 10 high       0      0
## # i 39,990 more rows
```

(b) With the coordinates of the form (seslow, sesmid) use the ses matrix and/or the information from

Question 6 to write down the coordinates of the ses levels "high", "mid" and "low" in terms of these

new variables.

low: (1,0) high : (0,0) mid : (0,1)

Q8. Since we are using general linear models, the model summary describes linear geometric objects, where the

dimension of the geometric object is determined by the number of continuous predictors. We have only a

single continuous predictor so our model describes a set of lines. How many lines are described by the model

in Question 5? Make sure you give some justification for your answer.

- **(Hint: see the Week 7 seminar and pages 2 – 6 of Module 7. The model summary and the ses matrix**

should help.)

solution : As there is only "age" which is continuous variable in our predictor and since we have "ses" variable with 3 levels(high,medium, low) and there are 4 other predictor variables with 2 levels(true, false), so ,we need a combination of 3 * 2 * 2* 2 * 2 = 48 lines to describe the model.

Q9. Now it is time to get serious with our data. There may be some interactions between the variables in the

data set, so fit a new model to your training set using all the individual variables and all the second-order

interaction terms. Use Anova() to find the p-values for each of the variables. Identify all interaction terms

that meet the 99.9% significance level.

```
surv_tn_logr2_ <- logistic_reg() %>% set_engine("glm")
surv_ft2_ <- surv_tn_logr2_ %>%
fit(recommend ~ .^2., data = surv_tn_)
summary(surv_ft2_$fit)
```

```
##
## Call:
## stats::glm(formula = recommend ~ .^2, family = stats::binomial,
```

```
##     data = data)
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -0.0306450  0.1968772  -0.156  0.87630
## age                               -0.0506298  0.0040580 -12.477  < 2e-16 ***
## company_awareTRUE                  0.2545222  0.1515537   1.679  0.09307 .
## malfunctionTRUE                   -1.7992030  1.5932031  -1.129  0.25877
## multi_purchTRUE                    3.1125331  0.1626080  19.141  < 2e-16 ***
## sesmid                            -0.0286535  0.1790067  -0.160  0.87283
## seslow                            -0.2972882  0.1820634  -1.633  0.10249
## social_mediaTRUE                   0.0986402  0.1580248   0.624  0.53249
## age:company_awareTRUE             -0.0036121  0.0028579  -1.264  0.20626
## age:malfunctionTRUE               -0.1165307  0.0371450  -3.137  0.00171 **
## age:multi_purchTRUE                0.0008269  0.0028563   0.289  0.77220
## age:sesmid                         0.0018485  0.0034212   0.540  0.58897
## age:seslow                         0.0080794  0.0034799   2.322  0.02025 *
## age:social_mediaTRUE              -0.0027619  0.0031124  -0.887  0.37488
## company_awareTRUE:malfunctionTRUE -0.3426335  0.4249302  -0.806  0.42005
## company_awareTRUE:multi_purchTRUE -0.0765064  0.0680594  -1.124  0.26097
## company_awareTRUE:sesmid          -0.0067087  0.0734750  -0.091  0.92725
## company_awareTRUE:seslow           0.0197427  0.0751176   0.263  0.79269
## company_awareTRUE:social_mediaTRUE -0.1870507  0.0830577  -2.252  0.02432 *
## malfunctionTRUE:multi_purchTRUE    0.1191206  0.7450590   0.160  0.87298
## malfunctionTRUE:sesmid            -0.0977084  0.5316806  -0.184  0.85419
## malfunctionTRUE:seslow            -0.5286499  0.5033113  -1.050  0.29356
## malfunctionTRUE:social_mediaTRUE  -0.6562694  0.8205893  -0.800  0.42385
## multi_purchTRUE:sesmid            -0.0813662  0.0788799  -1.032  0.30230
## multi_purchTRUE:seslow             0.4915649  0.0841754   5.840 5.23e-09 ***
## multi_purchTRUE:social_mediaTRUE  -0.0025667  0.0848956  -0.030  0.97588
## sesmid:social_mediaTRUE            0.0315412  0.0986667   0.320  0.74922
## seslow:social_mediaTRUE            0.2704105  0.1008547   2.681  0.00734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47800  on 39999  degrees of freedom
## Residual deviance: 29321  on 39972  degrees of freedom
## AIC: 29377
##
## Number of Fisher Scoring iterations: 10
```

```
#Anova
Anova(surv_ft2_$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: recommend
##                            LR Chisq Df Pr(>Chisq)
## age                          1487.8  1  < 2.2e-16 ***
## company_aware                   2.5  1  0.1121505
## malfunction                  5588.7  1  < 2.2e-16 ***
## multi_purch                 13085.8  1  < 2.2e-16 ***
```

```
## ses                           136.5  2  < 2.2e-16 ***
## social_media                    0.9  1  0.3453407
## age:company_aware               1.6  1  0.2066048
## age:malfunction                13.9  1  0.0001923 ***
## age:multi_purch                 0.1  1  0.7721921
## age:ses                         5.9  2  0.0524961 .
## age:social_media                0.8  1  0.3755808
## company_aware:malfunction       0.6  1  0.4226775
## company_aware:multi_purch       1.3  1  0.2604554
## company_aware:ses               0.1  2  0.9351341
## company_aware:social_media      5.1  1  0.0242100 *
## malfunction:multi_purch         0.0  1  0.8709613
## malfunction:ses                 1.3  2  0.5265058
## malfunction:social_media        0.6  1  0.4518418
## multi_purch:ses                54.8  2  1.257e-12 ***
## multi_purch:social_media        0.0  1  0.9758792
## ses:social_media                8.6  2  0.0134819 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q9. The interaction terms that meet the 99.9% significance are :

Here i am considering only age:malfunction and multi_purch:ses , as company_aware:social_m

Q10. We'll now apply backwards stepwise regression. As we learned in Module 7, best practice is to only remove

terms one-by-one starting with the least significant. However, our client wants a result ASAP, so we'll just

jump straight to removing all the interaction terms that are not extremely significant.

(a) So first fit a new model with just the individual variables and the significant interactions terms that you

identified in Question 9. Show the Anova() output.

```
surv_tn_ft2_ <- logistic_reg() %>% set_engine("glm") %>%
  fit(recommend ~ age + malfunction+ company_aware + multi_purch + ses + social_media + age:malfunction

#Anova
Anova(surv_tn_ft2_$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: recommend
##                 LR Chisq Df Pr(>Chisq)
## age               1485.7  1  < 2.2e-16 ***
## malfunction       5634.8  1  < 2.2e-16 ***
## company_aware        2.5  1  0.1133722
## multi_purch      13067.8  1  < 2.2e-16 ***
## ses                135.9  2  < 2.2e-16 ***
## social_media         0.9  1  0.3380740
## age:malfunction     13.6  1  0.0002217 ***
## multi_purch:ses     69.1  2  9.867e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

10 b.Then continue with the proper step-by-step backwards step-wise regression to find a model where all

terms (individual terms and interaction terms) meet the 95% significance level. At each step, identify

the variable that you will remove, and why you will choose that one. Then show the resulting Anova()

after you fit each model.

```
surv_tn_ft2_ <- logistic_reg() %>% set_engine("glm") %>%
  fit(recommend ~ age + malfunction+ company_aware + multi_purch + ses + age:malfunction + multi_purch:s

#Anova
Anova(surv_tn_ft2_$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: recommend
##                 LR Chisq Df Pr(>Chisq)
## age               1886.8  1  < 2.2e-16 ***
## malfunction       5635.1  1  < 2.2e-16 ***
## company_aware        2.5  1  0.1122812
## multi_purch      13067.2  1  < 2.2e-16 ***
## ses                136.1  2  < 2.2e-16 ***
## age:malfunction     13.7  1  0.0002186 ***
## multi_purch:ses     69.1  2  9.651e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
surv_tn_ft2_ <- logistic_reg() %>% set_engine("glm") %>%
  fit(recommend ~ age + malfunction + multi_purch + ses + age:malfunction + multi_purch:ses , data = su

#Anova
Anova(surv_tn_ft2_$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: recommend
##                 LR Chisq Df Pr(>Chisq)
## age               1886.9  1  < 2.2e-16 ***
## malfunction       5635.8  1  < 2.2e-16 ***
## multi_purch      13068.4  1  < 2.2e-16 ***
## ses                136.2  2  < 2.2e-16 ***
## age:malfunction     13.6  1  0.0002234 ***
## multi_purch:ses     69.1  2  9.815e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q11.

(a) Which interaction terms are significant in your final model?

solution: Here according to my dataset survey 1 age:malfunction and multi_purch:ses are the favourable conditions or parameters i have got as interaction terms which are significant in my model.

(b) Thinking about the context of the data, provide some reasonable hypotheses for why those interaction

terms might represent real effects (and are not just statistical noise).

solution: age:malfunction - So when considering this parameter , we usually tend to agree that most malfunctions are detected by younger generation rather than older generation, as the older generation don't have frequent usage towards gadgets and significance usage from them would be different as their usage towards gadgets may be of a minimum requirements.

multi_purch:ses -It represents more on socio-economic status and also the ability towards buying mu

```
summary(surv_tn_ft2_$fit)
```

```
##
## Call:
## stats::glm(formula = recommend ~ age + malfunction + multi_purch +
##     ses + age:malfunction + multi_purch:ses, family = stats::binomial,
```

```
##     data = data)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -0.034846   0.052765  -0.660 0.509001
## age                     -0.050490   0.001247 -40.478  < 2e-16 ***
## malfunctionTRUE         -3.060526   0.880038  -3.478 0.000506 ***
## multi_purchTRUE          3.078099   0.053389  57.654  < 2e-16 ***
## sesmid                   0.053705   0.046310   1.160 0.246176
## seslow                   0.205564   0.045548   4.513 6.39e-06 ***
## age:malfunctionTRUE     -0.104586   0.034258  -3.053 0.002266 **
## multi_purchTRUE:sesmid -0.066324   0.073430  -0.903 0.366402
## multi_purchTRUE:seslow  0.523250   0.077618   6.741 1.57e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47800  on 39999  degrees of freedom
## Residual deviance: 29347  on 39991  degrees of freedom
## AIC: 29365
##
## Number of Fisher Scoring iterations: 10
```

Q12. Write general form for logistic regression from your model.

Solution:

$$\hat{f}_i = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 malfunctionTRUE + \hat{\beta}_3 multipurchTRUE + \hat{\beta}_4 sesmid + \hat{\beta}_5 seslow + \hat{\beta}_6 age : malfunctionTRUE + \hat{\beta}_7 multi_p$$

where,

$\hat{f}$ = the estimated function.

$\hat{\beta}_0$ = estimated intercept.

$\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$, $\hat{\beta}_5$, $\hat{\beta}_6$,$\hat{\beta}_7$, $\hat{\beta}_8$ are the coefficients.

ϵ_i = error term.

Q13.Looking at Question 12, the geometric situation is slightly more complicated now than in Question 8, although

our model should still produce a set of lines.

```
Anova(surv_tn_ft2_$fit)
```

```
## Analysis of Deviance Table (Type II tests)
```

```
##
## Response: recommend
##                  LR Chisq Df Pr(>Chisq)
## age                1886.9  1  < 2.2e-16 ***
## malfunction        5635.8  1  < 2.2e-16 ***
## multi_purch       13068.4  1  < 2.2e-16 ***
## ses                 136.2  2  < 2.2e-16 ***
## age:malfunction      13.6  1  0.0002234 ***
## multi_purch:ses      69.1  2  9.815e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q13(a). How many lines does your final model describe? Make sure you provide some justification for your

answer.

Solution : So basically we have "age" where it is continuous variable in our predictors and also there is "ses" variable with 3 levels(high,medium and low) and two other predictor variables with two levels(true and false). So we need a combination of 3 * 2 * 2 = 12 lines to describe the model.

Q13(b). Are the lines all parallel? If not, explain why not.

Solution: The lines are not parallel from the nature of the model with anova and we can see that the significant p-values from it. There is strong relationship between the variables so the it is a non linear relationship.

Q14. Now output the summary of your final model showing the estimated coefficients, and use that to write ˆ fi

with all the estimated coefficients replacing the ˆ j pronumerals.

```
summary(surv_tn_ft2_$fit)
```

```
##
## Call:
## stats::glm(formula = recommend ~ age + malfunction + multi_purch +
##     ses + age:malfunction + multi_purch:ses, family = stats::binomial,
##     data = data)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -0.034846   0.052765  -0.660 0.509001
## age                    -0.050490   0.001247 -40.478  < 2e-16 ***
## malfunctionTRUE        -3.060526   0.880038  -3.478 0.000506 ***
## multi_purchTRUE         3.078099   0.053389  57.654  < 2e-16 ***
```

```
## sesmid                   0.053705   0.046310   1.160 0.246176
## seslow                   0.205564   0.045548   4.513 6.39e-06 ***
## age:malfunctionTRUE     -0.104586   0.034258  -3.053 0.002266 **
## multi_purchTRUE:sesmid -0.066324   0.073430  -0.903 0.366402
## multi_purchTRUE:seslow  0.523250   0.077618   6.741 1.57e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47800  on 39999  degrees of freedom
## Residual deviance: 29347  on 39991  degrees of freedom
## AIC: 29365
##
## Number of Fisher Scoring iterations: 10
```

$\hat{f}_i = -0.034846 + -0.050490 age + -3.060526 malfunctionTRUE + 3.078099 multi_purchTRUE + 0.053705 sesmid + 0.205564 sesl$

q15. What is our estimate for the log-odds for a respondent:

(a) who has a low Socio-Economic Status, yet purchased several Gadgets and none of them stopped working?

Solution:

$\hat{f}_i = -0.034846 + -0.050490 age + -3.060526 malfunctionTRUE + 3.078099 multi_purchTRUE + 0.053705 sesmid + 0.205564 sesl$

- valuea = -0.034846+ -0.050490 age + -3.060526 * 0 + 3.078099 * 1 + 0.053705 * 0 + 0.205564 * 1 + -0.104586 age * 0 +-0.066324 * 1 * 0 + 0.523250 * 1 *1
- = 3.772067 - 0.050490 age

(b) who has a mid-range Socio-Economic Status, only purchased a single Gadget and it broke?

- valueb = -0.034846+ -0.050490 age + -3.060526 * 1 + 3.078099 * 0 + 0.053705 * 1 + 0.205564 * 0 + -0.104586 age * 1 +-0.066324 * 0 * 1 + 0.523250 * 0 *0
- = - 3.04166 -0.155076 age

Q16.Now apply your final model to the testing data. Produce a new tibble containing the predicted class and the

prediction probabilities. Output the first 10 lines of this tibble.

```
pred_test_surv_ = predict(surv_tn_ft2_,
                new_data = surv_tt_,
                type = "class")

surv_pred_fin = predict(surv_tn_ft2_,
                      new_data= surv_tt_,
                      type = "prob") %>%
  bind_cols (surv_tt_ %>% select (recommend) ,
             pred_class = pred_test_surv_)

# Output the first 10 lines
head(surv_pred_fin,10)
```

```
## # A tibble: 10 x 4
##    .pred_no .pred_yes recommend .pred_class
##       <dbl>     <dbl> <fct>     <fct>
##  1    0.993 0.00683   no        no
##  2    0.776 0.224     no        no
##  3    0.925 0.0755    no        no
##  4    0.212 0.788     yes       yes
##  5    0.886 0.114     no        no
##  6    0.190 0.810     no        yes
##  7    0.376 0.624     yes       yes
##  8    0.267 0.733     yes       yes
##  9    1.00  0.0000110 no        no
## 10    0.140 0.860     yes       yes
```

Q17. Now we need to evaluate our model.

(a) Find the confusion matrix.

```
#confusion matrix
surv_pred_fin %>%  conf_mat( truth = recommend, estimate = .pred_class )
```

```
##           Truth
## Prediction   no  yes
##        no  6528 1008
##        yes  577 1887
```

(b) If leaving a review is classified as a success, find the sensitivity and specificity of our model.

```
#sensitivity (the ratio of correct positives to all positives)
my_sens=1887/(1008+1887)
my_sens
```

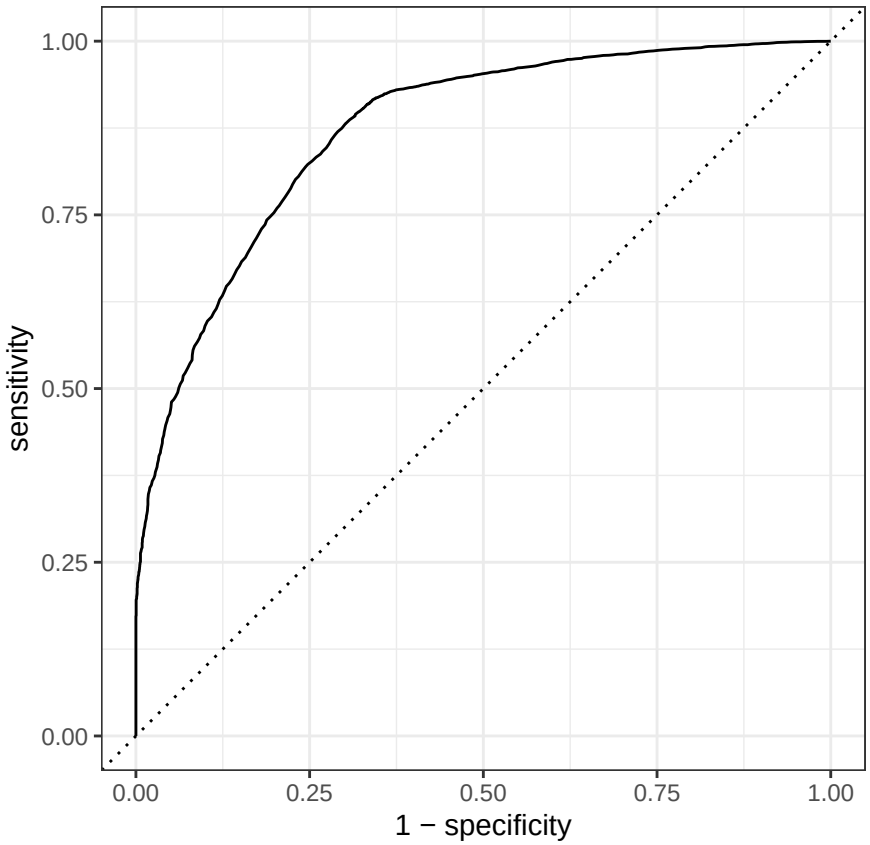
```
## [1] 0.6518135
```

```
#specificity (the ratio of the correct negatives to all negatives)
my_spec=6528/(6528+577)
my_spec
```

```
## [1] 0.9187896
```

(c) Plot the ROC curve.

```
surv_pred_fin %>%  roc_curve( .pred_no, truth = recommend) %>%
  autoplot()
```



(d) What is the AUC of this ROC curve?

```
auc = surv_pred_fin %>%
  roc_auc(.pred_no, truth = recommend)
auc
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.872
```

Q18. Finally, let's answer the company's question. Based on your model, do you predict that the Mayor will

recommend the Gadget 2? Write some text to interpret your results for the company, and make sure you give

the probabilities of your predicted class.

The AUC is 0.8718037 and the prediction for the model created above surv_tn_ft2_ is predicted for Mayor below , where the model will predict that the Mayor will recommend Gadget 2 . Also , the sucess of the model predicted is about 87.18 percentage.

```
mayor_predi_ <- predict(surv_tn_ft2_,
                        new_data = tibble(
                          age = 45 ,
                          malfunction = "FALSE" ,
                          multi_purch = "TRUE" ,
                          ses = "high"
                        ))

mayor_predi_
```

```
## # A tibble: 1 x 1
##   .pred_class
##   <fct>
## 1 yes
```