Gadget2 Analysis

Setup

```
#Load the required packages
library(tidyverse)
library(tidymodels)
pacman::p_load(modelr)
pacman::p_load(car)
```

Q1. Loading the data

```
# My student number goes here
ysn = 1908447
# Calculate my student number modulo 3
filenum <- (ysn + 2) %% 3
filenum
```

```
## [1] 2
```

```
filename <- paste0("./data/survey_",filenum,".csv")
filename
```

```
## [1] "./data/survey_2.csv"
```

```
# Read in the data
survey <- read_csv(filename)
# survey <- as_tibble(survey)
# Display the first 10 lines of the data
print(n=10,survey)
```

```
## # 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    49 TRUE          FALSE       FALSE       mid   TRUE
##  2         0    46 TRUE          TRUE        TRUE        low   FALSE
##  3         1    50 TRUE          FALSE       TRUE        mid   FALSE
##  4         0    37 FALSE         TRUE        FALSE       mid   TRUE
##  5         0    31 FALSE         FALSE       FALSE       high  TRUE
##  6         1    45 FALSE         FALSE       FALSE       low   FALSE
##  7         0    26 TRUE          TRUE        TRUE        mid   TRUE
##  8         0    29 TRUE          FALSE       FALSE       low   TRUE
##  9         0    64 FALSE         FALSE       FALSE       high  FALSE
## 10         1    47 TRUE          FALSE       FALSE       mid   TRUE
## # i 49,990 more rows
```

Q2. Identify what types of variables we now have

- recommend: Categorical Nominal
  - The "recommend" column is a feature that indicates whether a recommendation has been made or not, with a binary flag, and the values do not represent any order.
- age: Quantitative Continuous
  - The model created in this project accept the decimal values for "age". Therefore "age" is treated as decimal variable.
- company_aware: Categorical Nominal
  - The "company_aware" column contains TRUE or FALSE values, which are unordered categorical values.
- malfunction: Categorical Nominal
  - The "malfunction" column contains TRUE or FALSE values, which are unordered categorical values.
- multi_purch: Categorical Nominal
  - The "multi_purch" column contains TRUE or FALSE values, which are unordered categorical values.
- SES: Categorical Ordinal
  - The "malfunction" column contains "low","mid" and "high", which are ordered categorical values.
- social_media: Categorical Nominal
  - The "social_media" column contains TRUE or FALSE values, which are unordered categorical values.

Q3. Tame data

Make sure that all column names are in snake case

```r
# The snake case is a notation that connects lowercase words with an underscore.
# SES is in upper case, so it should be converted to lower case.
survey3_1 <- rename(survey,ses=SES)
```

Make the variables age, company_aware, malfunction, multi_purch and social_media conform to tame data

```r
print(n=10,survey)
```

```
## # 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    49 TRUE          FALSE       FALSE       mid   TRUE
##  2         0    46 TRUE          TRUE        TRUE        low   FALSE
##  3         1    50 TRUE          FALSE       TRUE        mid   FALSE
##  4         0    37 FALSE         TRUE        FALSE       mid   TRUE
##  5         0    31 FALSE         FALSE       FALSE       high  TRUE
##  6         1    45 FALSE         FALSE       FALSE       low   FALSE
##  7         0    26 TRUE          TRUE        TRUE        mid   TRUE
##  8         0    29 TRUE          FALSE       FALSE       low   TRUE
##  9         0    64 FALSE         FALSE       FALSE       high  FALSE
## 10         1    47 TRUE          FALSE       FALSE       mid   TRUE
## # i 49,990 more rows
```

The all columns, except for the "recommend" and "ses", meet the criteria for tame data.

Convert recommend to a fctor data type, with yes for 1 and no for 0.

```
survey3_2<-survey3_1
survey3_2$recommend <- survey3_2$recommend %>%
  as.factor() %>%
  fct_recode("yes" = "1", "no"="0")
```

Convert the Socio-Economic Status to a fctor.

```
survey3_3 <- survey3_2 %>%
  mutate(ses=factor(ses))

# Output the first 10 lines
print(n=10,survey3_3)
```

```
## # A tibble: 50,000 x 7
##    recommend   age company_aware malfunction multi_purch ses   social_media
##    <fct>     <dbl> <lgl>         <lgl>       <lgl>       <fct> <lgl>
##  1 no           49 TRUE          FALSE       FALSE       mid   TRUE
##  2 no           46 TRUE          TRUE        TRUE        low   FALSE
##  3 yes          50 TRUE          FALSE       TRUE        mid   FALSE
##  4 no           37 FALSE         TRUE        FALSE       mid   TRUE
##  5 no           31 FALSE         FALSE       FALSE       high  TRUE
##  6 yes          45 FALSE         FALSE       FALSE       low   FALSE
##  7 no           26 TRUE          TRUE        TRUE        mid   TRUE
##  8 no           29 TRUE          FALSE       FALSE       low   TRUE
##  9 no           64 FALSE         FALSE       FALSE       high  FALSE
## 10 yes          47 TRUE          FALSE       FALSE       mid   TRUE
## # i 49,990 more rows
```

Q4. Split data into a training set

```
# Set the seed
set.seed(ysn)
survey4 <- survey3_3

# Split data into a training set
survey_split <- initial_split( survey4, prop=4/5 ) # Create our split object
survey_train <- training( survey_split ) # Get our training sets
survey_test <- testing( survey_split ) # Get our testing sets

# Output the dimensions of our training and testing sets
dim(survey_train)
```

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

```
dim(survey_test)
```

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

Q5. Fit a logistic regression model to training data

```r
lr_spec <- logistic_reg(mode = "classification") %>%
  set_engine("glm")
survey_lr <- lr_spec %>%
  fit(recommend    ~ ., data = survey_train)

# Output the summary of the model
summary(survey_lr$fit)
```

```
##
## Call:
## stats::glm(formula = recommend ~ ., family = stats::binomial,
##     data = data)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.085843   0.075461   1.138    0.255
## age                 -0.052159   0.001391 -37.484   <2e-16 ***
## company_awareTRUE -0.030334   0.030046  -1.010    0.313
## malfunctionTRUE   -5.678819   0.183218 -30.995   <2e-16 ***
## multi_purchTRUE    3.185811   0.033000  96.539   <2e-16 ***
## seslow              0.341253   0.035875   9.512   <2e-16 ***
## sesmid             -0.001936   0.036684  -0.053    0.958
## social_mediaTRUE  -0.055858   0.040049  -1.395    0.163
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47975  on 39999  degrees of freedom
## Residual deviance: 29603  on 39992  degrees of freedom
## AIC: 29619
##
## Number of Fisher Scoring iterations: 8
```

Q6. See what happens to ses when a model is fitted

Q6(a). How many new variables have been introduced?

```r
model_matrix(survey_train,~ses)
```

```
## # A tibble: 40,000 x 3
##    `(Intercept)` seslow sesmid
##            <dbl>  <dbl>  <dbl>
##  1             1      0      1
##  2             1      0      0
##  3             1      1      0
##  4             1      0      1
##  5             1      0      0
##  6             1      0      1
##  7             1      0      0
##  8             1      1      0
```

```
## 9               1      0      1
## 10              1      0      0
## # i 39,990 more rows
```

The "ses" is decomposed into "seslow"("low"), "sesmid"("mid") and "intercept". In the model, two variables, "seslow" and "sesmid", are introduced.

Q6(b). What is the reference level for ses?

high("seshigh")

Q7(a). Build a new tibble called ses matrix

```r
ses_matrix <- survey_train$ses %>%
  cbind(select(model_matrix(survey_train,~ses),seslow,sesmid)) %>%
  as.tibble()
ses_matrix <- rename(ses_matrix, ses=.)
ses_matrix
```

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

Q7(b). Write down the coordinates of the ses levels

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

Q8. How many lines are described by the model in Q5?

The categorical predictors used by the model in Q5 are as follows:

- company_aware
- malfunction
- multi_purch
- seslow
- sesmid
- social_media

"company_aware", "malfunction", "multi_purch", and "social_media" have binary values, while "seslow" and "sesmid" are based on SES, thus resulting in three possible combinations of values, "low", "mid" and "high". Therefore, the number of lines is $2^4 \times 3 = 48$.

Q9. Fit a model to training set using all the individual variables and all the second-order interaction terms

```
# Fit a model to training set
survey_interact_lr <- lr_spec %>%
  fit(recommend    ~ .^2, data = survey_train)

# Find the p-values for each of the variables
Anova(survey_interact_lr$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: recommend
##                             LR Chisq Df Pr(>Chisq)
## age                           1563.3  1  < 2.2e-16 ***
## company_aware                    1.0  1   0.328114
## malfunction                   5742.5  1  < 2.2e-16 ***
## multi_purch                  12873.8  1  < 2.2e-16 ***
## ses                            121.8  2  < 2.2e-16 ***
## social_media                     1.6  1   0.201699
## age:company_aware                2.2  1   0.138926
## age:malfunction                 29.1  1  6.710e-08 ***
## age:multi_purch                  0.4  1   0.511581
## age:ses                          2.2  2   0.337691
## age:social_media                 2.3  1   0.128251
## company_aware:malfunction        4.3  1   0.038166 *
## company_aware:multi_purch        0.0  1   0.834211
## company_aware:ses                1.8  2   0.404090
## company_aware:social_media       1.0  1   0.312472
## malfunction:multi_purch          0.3  1   0.600582
## malfunction:ses                  0.9  2   0.625453
## malfunction:social_media         4.3  1   0.039081 *
## multi_purch:ses                 47.5  2  4.866e-11 ***
## multi_purch:social_media         7.8  1   0.005313 **
## ses:social_media                 0.5  2   0.780963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The following interaction terms meet the 99.9% significance level.

- age:malfunction
- multi_purch:ses

Q10. Apply backwards stepwise regression

Q10(a). Fit a new model with just the individual variables and the significant interactions terms

```
survey_lr_10a <- lr_spec %>%
  fit(recommend    ~ . + age:malfunction + multi_purch:ses, data = survey_train)

# Show the Anova() output.
Anova(survey_lr_10a$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: recommend
##                  LR Chisq Df Pr(>Chisq)
## age                1558.7  1  < 2.2e-16 ***
## company_aware         1.0  1     0.3112
## malfunction        5806.3  1  < 2.2e-16 ***
## multi_purch       12873.6  1  < 2.2e-16 ***
## ses                 122.2  2  < 2.2e-16 ***
## social_media          1.9  1     0.1722
## age:malfunction      22.0  1  2.724e-06 ***
## multi_purch:ses      56.5  2  5.422e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q10(b). Find a model where all terms meet the 95% significance level

```r
# Exclude the "company_aware", which has the highest p-value and does not meet 95% significance level
survey_lr_10b_1 <- lr_spec %>%
  fit(recommend    ~ age + malfunction + multi_purch + ses + social_media +
        age:malfunction + multi_purch:ses, data = survey_train)

Anova(survey_lr_10b_1$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: recommend
##                  LR Chisq Df Pr(>Chisq)
## age                1558.6  1  < 2.2e-16 ***
## malfunction        5805.6  1  < 2.2e-16 ***
## multi_purch       12873.2  1  < 2.2e-16 ***
## ses                 122.1  2  < 2.2e-16 ***
## social_media          1.9  1       0.17
## age:malfunction      22.0  1  2.747e-06 ***
## multi_purch:ses      56.5  2  5.391e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Exclude the "social_media", which has the highest p-value and does not meet 95% significance level
survey_lr_10b_2 <- lr_spec %>%
  fit(recommend    ~ age +  malfunction + multi_purch + ses +
        age:malfunction + multi_purch:ses, data = survey_train)

Anova(survey_lr_10b_2$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: recommend
##                  LR Chisq Df Pr(>Chisq)
## age                1959.9  1  < 2.2e-16 ***
## malfunction        5805.0  1  < 2.2e-16 ***
## multi_purch       12872.0  1  < 2.2e-16 ***
## ses                 122.1  2  < 2.2e-16 ***
## age:malfunction      22.0  1  2.668e-06 ***
## multi_purch:ses      56.5  2  5.330e-13 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All terms meet the 95% significance level now.

Q11. The significant interaction terms

Q11(a). Which interaction terms are significant in the final model?

According to the result of Q10(b), the following terms are significant interaction terms.

- age:malfunction
- multi_purch:ses

Q11(b). Provide some reasonable hypotheses for why those interaction terms might represent real effects

- age:malfunction

```r
survey_agg1 <- select(survey_train,age,malfunction)

survey_agg1<- survey_agg1 %>%
  mutate(age_10 = as.integer(paste0(str_replace(age,"\\d$",""), "0")),
         malfunction_flg = as.integer(ifelse(malfunction==TRUE, 1 ,0)) )

# The probability of malfunction per age group
survey_agg1 %>%
  group_by(age_10) %>%
  summarise(probability_malfunction=mean(malfunction_flg))
```

```
## # A tibble: 8 x 2
##   age_10 probability_malfunction
##    <int>                   <dbl>
## 1     10                   0.136
## 2     20                   0.156
## 3     30                   0.148
## 4     40                   0.147
## 5     50                   0.154
## 6     60                   0.147
## 7     70                   0.140
## 8     80                   0.194
```

```r
# The number of data points per age group
count(survey_agg1,age_10)
```

```
## # A tibble: 8 x 2
##   age_10     n
##    <int> <int>
## 1     10  1237
## 2     20 10692
## 3     30 10800
## 4     40  8671
## 5     50  5316
## 6     60  2497
## 7     70   720
## 8     80    67
```

Teenagers have fewer malfunctions, and there is a partial dependence between age and malfunctions. It can be presumed that the malfunction of teenagers is less likely to occur since fewer years have passed since the purchase of the gadget. On the other hand, the malfunction rate is particularly high in the 80s compared to other age groups. However, data points for individuals in their 80s are insufficient. Therefore, the malfunction rate of the 80s may be more susceptible to errors.

- multi_purch:ses

```
survey_agg2 <- select(survey_train,multi_purch,ses)
survey_agg2 <- survey_agg2 %>%
  mutate(multi_purch_flg = as.integer(ifelse(multi_purch==TRUE, 1 ,0)) )

# The probability of multiple purchases per SES
survey_agg2 %>%
  group_by(ses) %>%
  summarise(probability_multi_purch = mean(multi_purch_flg))
```

```
## # A tibble: 3 x 2
##   ses   probability_multi_purch
##   <fct>                   <dbl>
## 1 high                    0.300
## 2 low                     0.305
## 3 mid                     0.300
```

```
# The number of data points per SES
count(survey_agg2,ses)
```

```
## # A tibble: 3 x 2
##   ses       n
##   <fct> <int>
## 1 high  13380
## 2 low   13412
## 3 mid   13208
```

People who have a low Socio-Economic status tend to make multiple purchases compared to middle and high status, indicating a dependency between "multi_purch" and "ses". It can be speculated that people with a low Socio-Economic status may purchase multiple Gadgets, which are perceived as cool accessories, to maintain appearances. In contrast, individuals with mid or high status may not feel the need to show off. Therefore people with mid or high may not make multiple purchases.

Q12. Write down the general form of $\hat{f}_i$ for your final model in Question 10.

```
summary(survey_lr_10b_2$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)             0.059467   0.052345   1.136 0.255935
## age                    -0.051084   0.001243 -41.089  < 2e-16 ***
```

```
## malfunctionTRUE        -2.410610   0.808025  -2.983 0.002851 **
## multi_purchTRUE         2.997012   0.052836  56.723  < 2e-16 ***
## seslow                  0.163457   0.044687   3.658 0.000254 ***
## sesmid                 -0.025618   0.046255  -0.554 0.579688
## age:malfunctionTRUE    -0.121674   0.032203  -3.778 0.000158 ***
## multi_purchTRUE:seslow  0.528101   0.076482   6.905 5.03e-12 ***
## multi_purchTRUE:sesmid  0.062223   0.073567   0.846 0.397662
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47975  on 39999  degrees of freedom
## Residual deviance: 29528  on 39991  degrees of freedom
## AIC: 29546
##
## Number of Fisher Scoring iterations: 10
```

$$\hat{f}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{i}1} + \hat{\beta}_2 x_{\text{i}2} + \hat{\beta}_3 x_{\text{i}3} + \hat{\beta}_4 x_{\text{i}4} + \hat{\beta}_5 x_{\text{i}5} + \hat{\beta}_6 x_{\text{i}1} x_{\text{i}2} + \hat{\beta}_7 x_{\text{i}3} x_{\text{i}4} + \hat{\beta}_8 x_{\text{i}3} x_{\text{i}5}$$

- $\hat{f}_i$: Estimated function of the predictors
- $x_{\text{i}1}$: "age"
- $x_{\text{i}2}$(malfunctionTRUE): If "malfunction" is TRUE, $x_{\text{i}2}$ is equal to 1. Otherwise , $x_{\text{i}2}$ is equal to 0.
- $x_{\text{i}3}$(multi_purchTRUE): If "multi_purch" is TRUE, $x_{\text{i}3}$ is equal to 1. Otherwise , $x_{\text{i}3}$ is equal to 0.
- $x_{\text{i}4}$(seslow): If "ses" is "low", $x_{\text{i}4}$ is equal to 1. Otherwise , $x_{\text{i}4}$ is equal to 0.
- $x_{\text{i}5}$(sesmid): If "ses" is "mid", $x_{\text{i}5}$ is equal to 1. Otherwise , $x_{\text{i}5}$ is equal to 0.
- $\hat{\beta}_0$: Intercept
- $\hat{\beta}_1$: Coefficient of $x_{\text{i}1}$(age)
- $\hat{\beta}_2$: Coefficient of $x_{\text{i}2}$(malfunctionTRUE)
- $\hat{\beta}_3$: Coefficient of $x_{\text{i}3}$(multi_purchTRUE)
- $\hat{\beta}_4$: Coefficient of $x_{\text{i}4}$(seslow)
- $\hat{\beta}_5$: Coefficient of $x_{\text{i}5}$(sesmid)
- $\hat{\beta}_6$: Coefficient of interaction term between $x_{\text{i}1}$(age) and $x_{\text{i}2}$(malfunctionTRUE)
- $\hat{\beta}_7$: Coefficient of interaction term between $x_{\text{i}3}$(multi_purchTRUE) and $x_{\text{i}4}$(seslow)
- $\hat{\beta}_8$: Coefficient of interaction term between $x_{\text{i}3}$(multi_purchTRUE) and $x_{\text{i}5}$(sesmid)

Q13. The line of the final model

Q13(a). How many lines does your final model describe?

Values of interaction term depend on individual terms. Therefore the number of lines is counted based on individual categorical terms. The "malfunctionTRUE" and "multi_purchTRUE" have binary values, and while "seslow" and "sesmid" are based on "ses", thus resulting in three possible combinations of values, "low", "mid" and "high". Therefore, the number of lines is $2^2 \times 3 = 12$.

Q13(b). Are the lines all parallel?

No. One interaction term includes countinuous variable, $x_{\mathrm{i}1}$ ("age"). Therefore, the lines are not all parallel. For example, suppose there are two lines, $\hat{y}_0$ and $\hat{y}_1$. $\hat{y}_0$ corresponds to x_2 ("malfunctionTRUE") being 0, while $\hat{y}_1$ corresponds to x_2 being 1. To think simply, The values of other categorical terms are assumed to be 0. From the equation in Q12, $\hat{y}_1 - \hat{y}_0$ becomes as follows.

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{\mathrm{i}1}$$

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_{\mathrm{i}1} + \hat{\beta}_2 + \hat{\beta}_6 x_{\mathrm{i}1}$$

$$\hat{y}_1 - \hat{y}_0 = \hat{\beta}_2 + \hat{\beta}_6 x_{\mathrm{i}1}$$

From the above example, some lines are not parallel.

- $\hat{y}_0$: Estimated function of the predictors where all categorical terms are 0
- $\hat{y}_1$: Estimated function of the predictors where x_2 ("malfunctionTRUE") is eqaul to 1 and the rest of the other categorical terms are 0
- $x_{\mathrm{i}1}$: "age"
- $x_{\mathrm{i}2}$(malfunctionTRUE): If "malfunction" is TRUE, $x_{\mathrm{i}2}$ is equal to 1. Otherwise , $x_{\mathrm{i}2}$ is equal to 0. However $x_{\mathrm{i}2}$ is treated as 0 or 1 in Q13(b)
- $\hat{\beta}_0$: Intercept
- $\hat{\beta}_1$: Coefficient of $x_{\mathrm{i}1}$(age)
- $\hat{\beta}_2$: Coefficient of $x_{\mathrm{i}2}$(malfunctionTRUE)
- $\hat{\beta}_6$: Coefficient of interaction term between $x_{\mathrm{i}1}$(age) and $x_{\mathrm{i}2}$(malfunctionTRUE)

Q14. Output the summary of the final model and write log-odds with all the estimated coefficients

```r
# Output the summary of the final model
summary(survey_lr_10b_2$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)             0.059467   0.052345   1.136 0.255935
## age                    -0.051084   0.001243 -41.089  < 2e-16 ***
## malfunctionTRUE        -2.410610   0.808025  -2.983 0.002851 **
## multi_purchTRUE         2.997012   0.052836  56.723  < 2e-16 ***
## seslow                  0.163457   0.044687   3.658 0.000254 ***
## sesmid                 -0.025618   0.046255  -0.554 0.579688
## age:malfunctionTRUE    -0.121674   0.032203  -3.778 0.000158 ***
## multi_purchTRUE:seslow  0.528101   0.076482   6.905 5.03e-12 ***
## multi_purchTRUE:sesmid  0.062223   0.073567   0.846 0.397662
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47975  on 39999  degrees of freedom
## Residual deviance: 29528  on 39991  degrees of freedom
## AIC: 29546
##
## Number of Fisher Scoring iterations: 10
```

$\hat{f}_i = 0.0595 - 0.0511x_{\text{i1}} - 2.41x_{\text{i2}} + 3.00x_{\text{i3}} + 0.163x_{\text{i4}} - 0.0256x_{\text{i5}} - 0.122x_{\text{i1}}x_{\text{i2}} + 0.528x_{\text{i3}}x_{\text{i4}} + 0.0622x_{\text{i3}}x_{\text{i5}}$

- $\hat{f}_i$: Estimated function of the predictors
- x_{i1}: "age"
- x_{i2}(malfunctionTRUE): If "malfunction" is TRUE, x_{i2} is equal to 1. Otherwise , x_{i2} is equal to 0.
- x_{i3}(multi_purchTRUE): If "multi_purch" is TRUE, x_{i3} is equal to 1. Otherwise , x_{i3} is equal to 0.
- x_{i4}(seslow): If "ses" is "low", x_{i4} is equal to 1. Otherwise , x_{i4} is equal to 0.
- x_{i5}(sesmid): If "ses" is "mid", x_{i5} is equal to 1. Otherwise , x_{i5} is equal to 0.

Q15. What is the estimate for the log-odds for a respondent?

Q15(a). who has a low Socio-Economic Status, yet purchased several Gadgets and none of them stopped working?

```r
q15a_coef = survey_lr_10b_2$fit$coefficients

# Calculate total number of the intercept
q15a_coef["(Intercept)"] +
q15a_coef["malfunctionTRUE"]*0 +
q15a_coef["multi_purchTRUE"]*1 +
q15a_coef["seslow"]*1 +
q15a_coef["sesmid"]*0 +
q15a_coef["multi_purchTRUE:seslow"]*1*1 +
q15a_coef["multi_purchTRUE:sesmid"]*1*0
```

```
## (Intercept)
##    3.748037
```

```r
# Calculate the slope of the age
q15a_coef["age"] + q15a_coef["age:malfunctionTRUE"]*0
```

```
##         age
## -0.05108434
```

$$\hat{f}_i = 3.75 - 0.0511x_{\text{i1}}$$

- $\hat{f}_i$: Estimated function of the predictors
- x_{i1}: "age"

Q15(b). Who has a mid-range Socio-Economic Status, only purchased a single Gadget and it broke?

```
q15b_coef = survey_lr_10b_2$fit$coefficients

# Calculate total number of the intercept
q15b_coef["(Intercept)"] +
q15b_coef["malfunctionTRUE"]*1 +
q15b_coef["multi_purchTRUE"]*0 +
q15b_coef["seslow"]*0 +
q15b_coef["sesmid"]*1 +
q15b_coef["multi_purchTRUE:seslow"]*0*0 +
q15b_coef["multi_purchTRUE:sesmid"]*0*1
```

```
## (Intercept)
##   -2.376761
```

```
# Calculate total number of the slope
q15b_coef["age"]+q15b_coef["age:malfunctionTRUE"]*1
```

```
##        age
## -0.1727581
```

$$\hat{f}_i = -2.38 - 0.173x_{i1}$$

- $\hat{f}_i$: Estimated function of the predictors
- x_{i1}: "age"

Q16. Apply your final model to the testing data

```
# prediction probabilities.
prediction16 <- predict(survey_lr_10b_2,survey_test,  type= "prob")

prediction16 <- prediction16 %>%
  cbind(predict(survey_lr_10b_2,survey_test,  type= "class")  ) %>%
  as.tibble()

# Output the first 10 lines
print(n=10,prediction16)
```

```
## # A tibble: 10,000 x 3
##    .pred_no .pred_yes .pred_class
##       <dbl>     <dbl> <fct>
##  1    0.821  0.179    no
##  2    0.997  0.00297  no
##  3    0.999  0.00106  no
##  4    0.292  0.708    yes
##  5    0.153  0.847    yes
##  6    0.999  0.000952 no
##  7    0.893  0.107    no
##  8    0.878  0.122    no
##  9    0.164  0.836    yes
## 10    0.914  0.0857   no
## # i 9,990 more rows
```

Q17. Evaluate our model

Q17(a). Find the confusion matrix.

```
prediction17<-prediction16
prediction17$ground_truth <- survey_test$recommend

prediction17 %>%
  conf_mat( truth = ground_truth, estimate = .pred_class )
```

```
##           Truth
## Prediction   no  yes
##        no  6597  965
##        yes  502 1936
```

Q17(b). If leaving a review is classified as a success, find the sensitivity and specificity of our model.

```
# Display sensitivity
sensitivity <- 6597  / (6597 + 502)
sensitivity
```

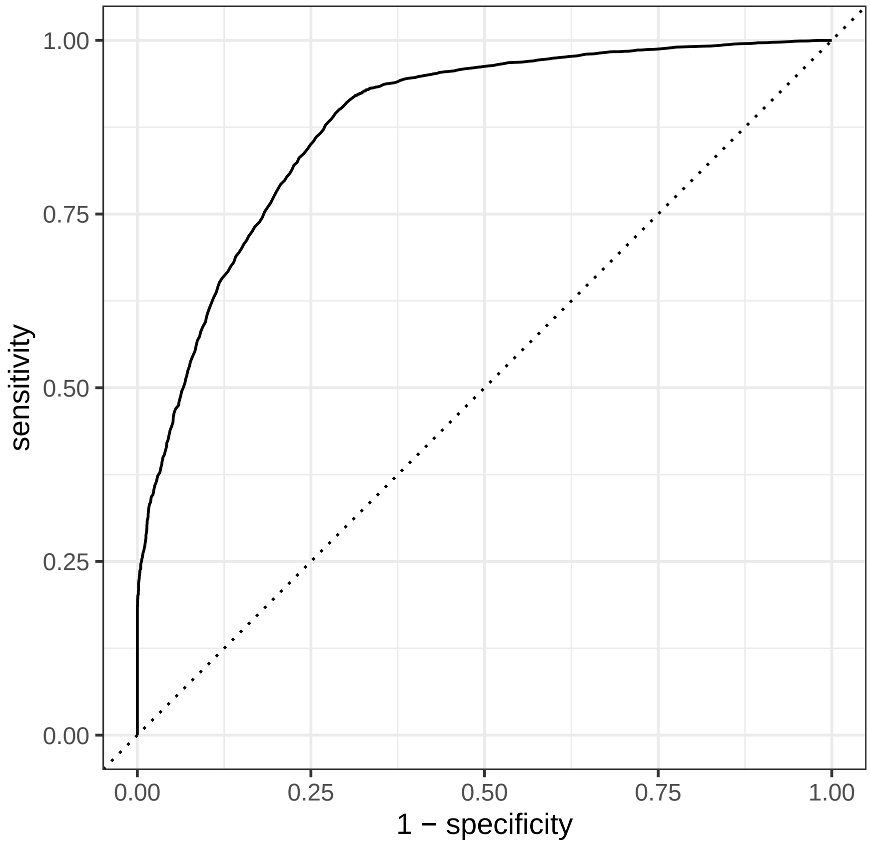
```
## [1] 0.9292858
```

```
# Display specificity
specificity <- 1936 / (1936 + 965)
specificity
```

```
## [1] 0.6673561
```

Q17(c). Plot the ROC curve.

```
prediction17 %>%
  roc_curve(.pred_no, truth = ground_truth) %>%
  autoplot()
```



Q17(d). What is the AUC of this ROC curve?

```
prediction17 %>%
  roc_auc(.pred_no, truth = ground_truth)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.880
```

Q18. Predict that the Mayor will recommend the Gadget2?

To summarise the mayor's information, His age is 45. Guessing from wearing a jacket with the company's logo on it, he knows the name of the company. Also, He purchased more than one Gadget and don't have any Gadgets® malfunctioned. Futhermore, his Socio-Economic Status is high, and He is an active social media user.

```
mayor_data  <-
  tibble(
    age   = 45,
    company_aware = TRUE,
    malfunction = FALSE,
    multi_purch = TRUE,
    ses = "high",
    social_media = TRUE
  )
```

```r
predict(survey_lr_10b_2,mayor_data,  type= "prob")
```

```
## # A tibble: 1 x 2
##   .pred_no .pred_yes
##      <dbl>     <dbl>
## 1    0.319     0.681
```

Based on the mayor's information and the prediction, the mayor recommends Gadget 2® with a probability of 68.1%.