HW8_Social_Media_Logistics_Regression_Yuefei_Chen

Yuefei Chen

2024-04-20

Model development

Running the following code, we build a multiple regression model based on rent house data. Its independent variables "Instagram_value", "Linkedin_value", "Snapchat_value", "Twitter_value", "Whatsapp_Wechat_value", "Youtube_value", "OTT_Netflix_Hulu_Prime_video_value", "Reddit_value", "job_interview_calls", "networking_done_with_coffee_chats", "learning_done_in_terms_of_items_created". The dependent variable is "Tired_waking_up_in_morning".

```
library(readr)
library(ggplot2)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
APP_data <- read_csv("Dataset/Social Media_cleaned.csv")
```

```
## New names:
## * `Hours_spent` -> `Hours_spent...3`
## * `Hours_spent` -> `Hours_spent...6`
## * `Hours_spent` -> `Hours_spent...9`
## * `Hours_spent` -> `Hours_spent...15`
## * `Hours_spent` -> `Hours_spent...18`
## * `Hours_spent` -> `Hours_spent...21`
## * `Hours_spent` -> `Hours_spent...24`

## Rows: 23 Columns: 33
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (15): ID, Instagram, Linkedin, Snapchat, Twitter, Whatsapp_Wechat, Yout...
## dbl  (12): Instagram_value, Linkedin_value, Snapchat_value, Twitter_value, W...
## time  (6): Hours_spent...3, Hours_spent...6, Hours_spent...9, Hours spent, H...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
APP_data <- APP_data[c(1:22), c(1:2, 4:5, 7:8, 10:11, 13:14, 16:17, 19:20, 22:23, 25:33)]
str(APP_data)
```

```
## tibble [22 x 25] (S3: tbl_df/tbl/data.frame)
##  $ ID                                       : chr [1:22] "masinl" "peace" "Patty" "Bunny" ...
##  $ Instagram                                : chr [1:22] "Yes" "Yes" "Yes" "Yes" ...
##  $ Instagram_value                          : num [1:22] 3.5 7.73 3.77 5.38 0 2.33 5.37 7 8.65 0.17
##  $ Linkedin                                 : chr [1:22] "Yes" "Yes" "Yes" "Yes" ...
##  $ Linkedin_value                           : num [1:22] 4 5.2 7 5.32 0.58 7 4 4 10 0 ...
##  $ Snapchat                                 : chr [1:22] "Yes" "Yes" "Yes" "Yes" ...
##  $ Snapchat_value                           : num [1:22] 1 3.68 0.53 1.3 0 0.47 0 3 3.83 0 ...
##  $ Twitter                                  : chr [1:22] "Yes" "No" "No" "No" ...
##  $ Twitter_value                            : num [1:22] 5 0 0 0 0.67 0 0 0 0 0 ...
##  $ Whatsapp_Wechat                          : chr [1:22] "Yes" "Yes" "Yes" "Yes" ...
##  $ Whatsapp_Wechat_value                    : num [1:22] 1 4.18 9.83 5.3 3 12 6 10 6.15 1 ...
##  $ Youtube                                  : chr [1:22] "Yes" "Yes" "Yes" "Yes" ...
##  $ Youtube_value                            : num [1:22] 2.5 4.25 1.85 2 3.5 7 3 2 4 3 ...
##  $ OTT_Netflix_Hulu_Prime video             : chr [1:22] "Yes" "No" "Yes" "Yes" ...
##  $ OTT_Netflix_Hulu_Prime_video_value       : num [1:22] 14.5 0 2 2 2 3 0 3 3 0 ...
##  $ Reddit                                   : chr [1:22] "Yes" "No" "No" "No" ...
##  $ Reddit_value                             : num [1:22] 2.5 0 0 0 1 0 0 0 0 0 ...
##  $ Application_type_Social media_OTT_Learning: chr [1:22] "OTT" "Social Media" "Social Media" "Social
##  $ job_interview_calls                      : num [1:22] 0 0 0 2 0 0 0 0 1 0 ...
##  $ networking_done_with_coffee_chats        : num [1:22] 0 1 0 0 2 0 2 0 0 0 ...
##  $ learning_done_in_terms_of_items_created  : num [1:22] 3 3 4 4 4 4 3 2 6 2 ...
##  $ Mood_Productivity                        : chr [1:22] "Yes" "Yes" "Yes" "Yes" ...
##  $ Tired_waking_up_in_morning               : chr [1:22] "No" "No" "No" "No" ...
##  $ Trouble_falling_asleep                   : chr [1:22] "No" "Yes" "No" "No" ...
##  $ felt_the_entire_week                     : num [1:22] 3 3 4 4 3 5 4 4 3 2 ...
```

```r
APP_data$Tired_waking_up_in_morning <- as.factor(APP_data$Tired_waking_up_in_morning)
#reg_data$area <- as.factor(reg_data$area)
#reg_data$rooms <- as.factor(reg_data$rooms)
#reg_data$bathroom <- as.factor(reg_data$rooms)
#reg_data$rent_amount <- as.factor(reg_data$rent_amount)
logistic <- glm(Tired_waking_up_in_morning~Instagram_value + Linkedin_value + Snapchat_value + Twitter_
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Model Acceptance and Residual Analysis

In the summary of the model, we focus on R squared value, coefficients, and P-value of each coefficient. The R-squared value is 1 and p value is 0.003830217. The Pseudo R-square shows there is an improvement and better than baseline model. The p-value based on Likelihood Ratio Test shows that the model improvement is significant. The AIC value is 25.

```r
summary(logistic)
```

```
##
## Call:
## glm(formula = Tired_waking_up_in_morning ~ Instagram_value +
##     Linkedin_value + Snapchat_value + Twitter_value + Whatsapp_Wechat_value +
##     Youtube_value + OTT_Netflix_Hulu_Prime_video_value + Reddit_value +
##     job_interview_calls + networking_done_with_coffee_chats +
##     learning_done_in_terms_of_items_created, family = "binomial",
##     data = APP_data)
##
## Coefficients:
##                                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                 149.87  224245.98   0.001    0.999
## Instagram_value                              81.67   54138.45   0.002    0.999
## Linkedin_value                              -83.49   37871.65  -0.002    0.998
## Snapchat_value                              -55.48   69825.80  -0.001    0.999
## Twitter_value                              -733.98  249046.93  -0.003    0.998
## Whatsapp_Wechat_value                       -44.65   39967.73  -0.001    0.999
## Youtube_value                              -170.27   75162.26  -0.002    0.998
## OTT_Netflix_Hulu_Prime_video_value           71.64   45451.68   0.002    0.999
## Reddit_value                                 32.48   22129.84   0.001    0.999
## job_interview_calls                        -201.82  105010.15  -0.002    0.998
## networking_done_with_coffee_chats           112.53   80398.13   0.001    0.999
## learning_done_in_terms_of_items_created     185.57   69364.98   0.003    0.998
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.7522e+01  on 21  degrees of freedom
## Residual deviance: 1.7479e-08  on 10  degrees of freedom
## AIC: 24
##
## Number of Fisher Scoring iterations: 25
```

```
ll.null <- logistic$null.deviance/-2
ll.proposed <- logistic$deviance/-2
(ll.null - ll.proposed) / ll.null
```

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

```
1 - pchisq(2*(ll.proposed - ll.null), df=(length(logistic$coefficients)-1))
```

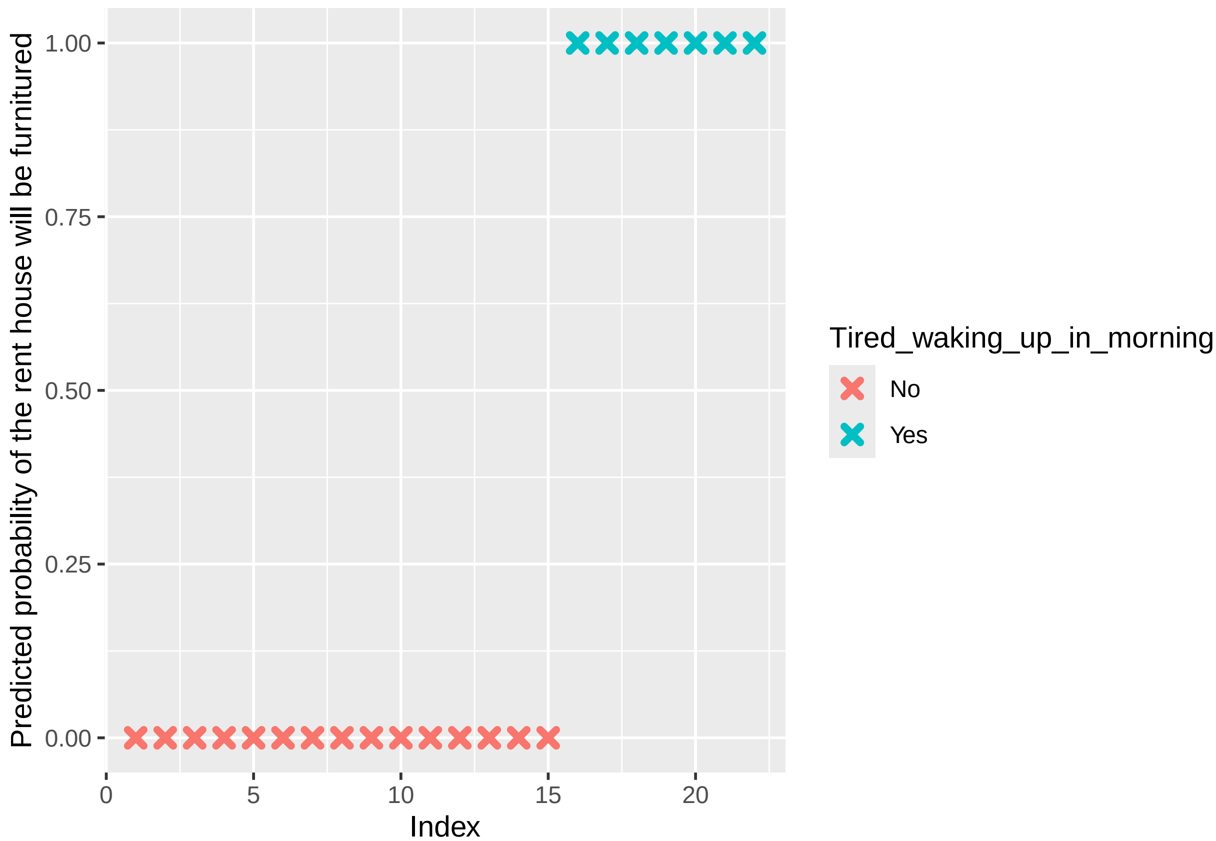
```
## [1] 0.003830217
```

Prediction

The data will be predicted in the model and the predicted probability of the rent house will be furnitured table is as follows.

```
predicted.data <- data.frame(probability.of.Tired_waking_up_in_morning=logistic$fitted.values,Tired_waki
predicted.data <- predicted.data[order(predicted.data$probability.of.Tired_waking_up_in_morning, decreas
predicted.data$rank <- 1:nrow(predicted.data)
ggplot(data=predicted.data, aes(x=rank, y=probability.of.Tired_waking_up_in_morning)) +
geom_point(aes(color=Tired_waking_up_in_morning), alpha=1, shape=4, stroke=2) +
```

```r
xlab("Index") +
ylab("Predicted probability of the rent house will be furnitured")
```



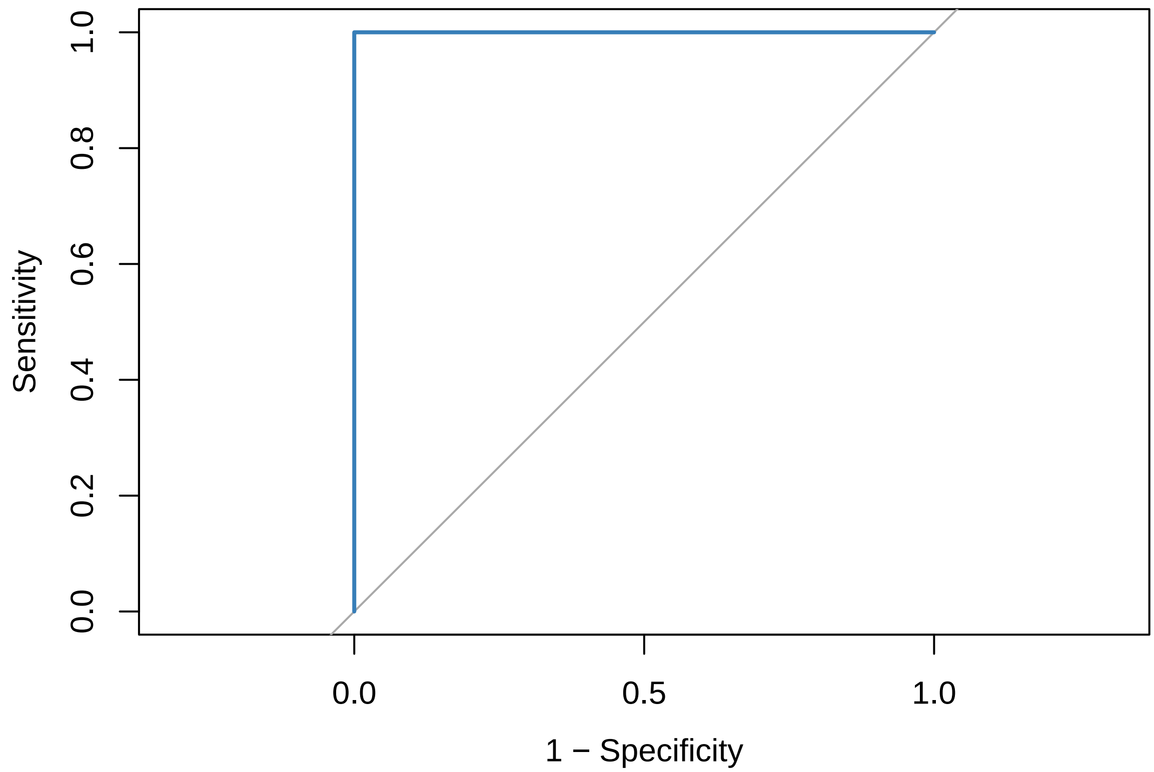
Model Accuracy

Receiver operating characteristic (ROC) curves is shown as follows. The area under the curve (AUC) is 1, which means the accuracy of the model is considered outstanding.

```r
roc(APP_data$Tired_waking_up_in_morning,logistic$fitted.values,plot=TRUE, legacy.axes=TRUE, col="#377eb8
```

```
## Setting levels: control = No, case = Yes

## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = APP_data$Tired_waking_up_in_morning, predictor = logistic$fitted.values,     p
##
## Data: logistic$fitted.values in 15 controls (APP_data$Tired_waking_up_in_morning No) < 7 cases (APP_d
## Area under the curve: 1
```