DS501 – Case Study 3: Mental Health & Social Media Balance

Corrin Courville

2025-12-06

Introduction

In this current report, the linkage between the state of mental well-being and social media usage is examined through the *Mental Health Social Media Balance* Dataset on Kaggle. The overall intention here is to explain the linkage between habit-driven activities such as screen time, sleep quality, stress levels, exercise levels, and social media platform preferences and happiness levels. The overall methodology here involves data preprocessing, creating a binary outcome variable, building a logistic regression model, model evaluation, visualization, and the creation of an interactive Shiny application. The application allows one to explore model-based outputs and make new predictions.

Dataset Description

The Mental Health Social Media Balance Dataset is the dataset used for this research. The Mental Health Social Media Balance Dataset is composed of self-reported behavior and lifestyle factors for social media participants. The various factors used here include age, gender, amount of time spent on screens each day, Sleep Quality (rate from 1-10), Stress Level (rate from 1-10), exercise undertaken each week, number of days without social media access, favorite social media platform, and Happiness Index (rate from 1-10).

The data was imported from the project’s data directory. This is the structure and content that defines the data.

```r
data_path <- "data/Mental Health Social Media Balance Dataset.csv"
raw_df <- read.csv(data_path, stringsAsFactors = FALSE)
str(raw_df)
```

```
## 'data.frame':    500 obs. of  10 variables:
##  $ User_ID                  : chr  "U001" "U002" "U003" "U004" ...
##  $ Age                      : int  44 30 23 36 34 38 26 26 39 39 ...
##  $ Gender                   : chr  "Male" "Other" "Other" "Female" ...
##  $ Daily_Screen_Time.hrs.   : num  3.1 5.1 7.4 5.7 7 6.6 7.8 7.4 4.7 6.6 ...
##  $ Sleep_Quality.1.10.      : num  7 7 6 7 4 5 4 5 7 6 ...
##  $ Stress_Level.1.10.       : num  6 8 7 8 7 7 8 6 7 8 ...
##  $ Days_Without_Social_Media: num  2 5 1 1 5 4 2 1 6 0 ...
##  $ Exercise_Frequency.week. : num  5 3 3 1 1 3 0 4 1 2 ...
##  $ Social_Media_Platform    : chr  "Facebook" "LinkedIn" "YouTube" "TikTok" ...
##  $ Happiness_Index.1.10.    : num  10 10 6 8 8 8 7 7 9 7 ...
```

Motivation

The rise of social media has raised concerns about the effect that online behavior can have on mental health. In recognition of the amount of time spent online and known correlations between sleep, physical activity, stress, and happiness, this research uses statistical analysis and interactive visualization to help shed light on this issue. In creating a model and implementing a Shiny application associated with the model, this research offers a means for examining happiness indicators that correspond with lifestyle decisions.

The fully deployed Shiny app is available here: https://cmcourville.shinyapps.io/ds501-hw6/

The tool provides the ability to analyze model output, evaluate prediction metrics, and perform a new case simulation. It represents the full process of doing data science from preparation to deployment.

Data Preparation

Before building the model, some preprocessing was done. The column names with dots and specials were handled for standardization. Categorical columns such as gender and platforms were transformed to factors. A significant transformation was carried out for creating a target variable with a range of values as Low_Happiness = Happiness_Index <=5. It can be observed that a simple classifier can be used on this target due to its binary nature.

```
df <- raw_df
df <- rename(
  df,
  Daily_Screen_Time_hrs = Daily_Screen_Time.hrs.,
  Sleep_Quality         = Sleep_Quality.1.10.,
  Stress_Level          = Stress_Level.1.10.,
  Exercise_Frequency    = Exercise_Frequency.week.,
  Happiness_Index       = Happiness_Index.1.10.
)
df <- mutate(
  df,
  Gender                = factor(Gender),
  Social_Media_Platform = factor(Social_Media_Platform)
)
```

The `Low_Happiness` variable was constructed as follows.

```
df <- df %>%
  mutate(
    Low_Happiness = ifelse(Happiness_Index <= 5, 1, 0)
  ) %>%
  select(
    Low_Happiness,
    Age,
    Gender,
    Daily_Screen_Time_hrs,
    Sleep_Quality,
    Stress_Level,
    Days_Without_Social_Media,
    Exercise_Frequency,
    Social_Media_Platform,
    Happiness_Index
  ) %>%
  na.omit()

summary(df$Low_Happiness)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.046   0.000   1.000
```

Machine Learning Methodology

Logistic regression was selected based on the fact that the response variable is binary. In this case, the response variable is `Low_Happiness`. Logistic regression predicts the probability that a subject is from the response category given the other predictor variables. The resulting model is interpretable and provides information on the contribution made by each predictor variable to determine the subjects with low happiness. It is made up of a linear predictor and the logistics function.

The predictors included in the model were age, gender, daily screen time, sleep quality, stress level, days without social media, exercise frequency, and social media platform. These predictors were chosen because they represent common lifestyle factors that plausibly influence well-being.

```
logit_formula <- Low_Happiness ~ Age + Gender + Daily_Screen_Time_hrs +
  Sleep_Quality + Stress_Level + Days_Without_Social_Media +
  Exercise_Frequency + Social_Media_Platform

logit_model <- glm(logit_formula, data = df, family = binomial())
summary(logit_model)
```

```
##
## Call:
## glm(formula = logit_formula, family = binomial(), data = df)
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -7.50513    3.99451  -1.879  0.06026 .
## Age                               -0.03748    0.03603  -1.040  0.29826
## GenderMale                        -0.52228    0.72685  -0.719  0.47242
## GenderOther                      -17.60056 1475.30491  -0.012  0.99048
## Daily_Screen_Time_hrs              0.33315    0.32561   1.023  0.30624
## Sleep_Quality                     -1.95701    0.47241  -4.143 3.43e-05 ***
## Stress_Level                       1.08025    0.38574   2.800  0.00510 **
## Days_Without_Social_Media          0.10471    0.17296   0.605  0.54490
## Exercise_Frequency                 0.83771    0.34360   2.438  0.01477 *
## Social_Media_PlatformInstagram     1.48953    1.23094   1.210  0.22625
## Social_Media_PlatformLinkedIn      2.32019    1.31996   1.758  0.07879 .
## Social_Media_PlatformTikTok        1.76937    1.24618   1.420  0.15566
## Social_Media_PlatformX (Twitter)  -1.19987    1.87899  -0.639  0.52310
## Social_Media_PlatformYouTube       3.46255    1.30674   2.650  0.00805 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 186.565  on 499  degrees of freedom
## Residual deviance:  68.057  on 486  degrees of freedom
## AIC: 96.057
##
## Number of Fisher Scoring iterations: 18
```

Mathematical Background

The logistic regression model estimates the probability that an observation belongs to the positive class (here, `Low_Happiness = 1`) using the logistic function:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$$

The expression inside the exponential term,

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

is called the **linear predictor**. Logistic regression maps this linear predictor into a probability between zero and one. Each predictor coefficient β_j describes how that variable influences the log-odds of low happiness, and exponentiating the coefficient gives an **odds ratio**:

$$\text{Odds Ratio} = e^{\beta_j}$$

```
logit_formula <- Low_Happiness ~ Age + Gender + Daily_Screen_Time_hrs +
  Sleep_Quality + Stress_Level + Days_Without_Social_Media +
  Exercise_Frequency + Social_Media_Platform

logit_model <- glm(logit_formula, data = df, family = binomial())
summary(logit_model)
```

```
##
## Call:
## glm(formula = logit_formula, family = binomial(), data = df)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -7.50513    3.99451  -1.879  0.06026 .
## Age                               -0.03748    0.03603  -1.040  0.29826
## GenderMale                        -0.52228    0.72685  -0.719  0.47242
## GenderOther                      -17.60056 1475.30491  -0.012  0.99048
## Daily_Screen_Time_hrs              0.33315    0.32561   1.023  0.30624
## Sleep_Quality                     -1.95701    0.47241  -4.143 3.43e-05 ***
## Stress_Level                       1.08025    0.38574   2.800  0.00510 **
## Days_Without_Social_Media          0.10471    0.17296   0.605  0.54490
## Exercise_Frequency                 0.83771    0.34360   2.438  0.01477 *
## Social_Media_PlatformInstagram     1.48953    1.23094   1.210  0.22625
## Social_Media_PlatformLinkedIn      2.32019    1.31996   1.758  0.07879 .
## Social_Media_PlatformTikTok        1.76937    1.24618   1.420  0.15566
## Social_Media_PlatformX (Twitter)  -1.19987    1.87899  -0.639  0.52310
## Social_Media_PlatformYouTube       3.46255    1.30674   2.650  0.00805 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 186.565  on 499  degrees of freedom
## Residual deviance:  68.057  on 486  degrees of freedom
## AIC: 96.057
##
## Number of Fisher Scoring iterations: 18
```

Model Evaluation

Predicted probabilities were created and discretized at a threshold of 0.5 for classification as Low_Happiness and otherwise. A confusion matrix was calculated for evaluating the classification accuracy, with parameters such as accuracy, sensitivity, and specificity.

```
probs <- predict(logit_model, type = "response")
pred <- ifelse(probs >= 0.5, 1, 0)
cm <- table(Predicted = pred, Actual = df$Low_Happiness)
cm
```

```
##          Actual
## Predicted   0   1
##         0 473  11
##         1   4  12
```

Accuracy, sensitivity, and specificity were calculated to summarize how well the model distinguishes low-happiness cases.

```
tp <- cm["1", "1"]
tn <- cm["0", "0"]
fp <- cm["1", "0"]
fn <- cm["0", "1"]

accuracy    <- (tp + tn) / sum(cm)
sensitivity <- tp / (tp + fn)
specificity <- tn / (tn + fp)

data.frame(
  Metric = c("Accuracy", "Sensitivity", "Specificity"),
  Value  = round(c(accuracy, sensitivity, specificity), 3)
)
```

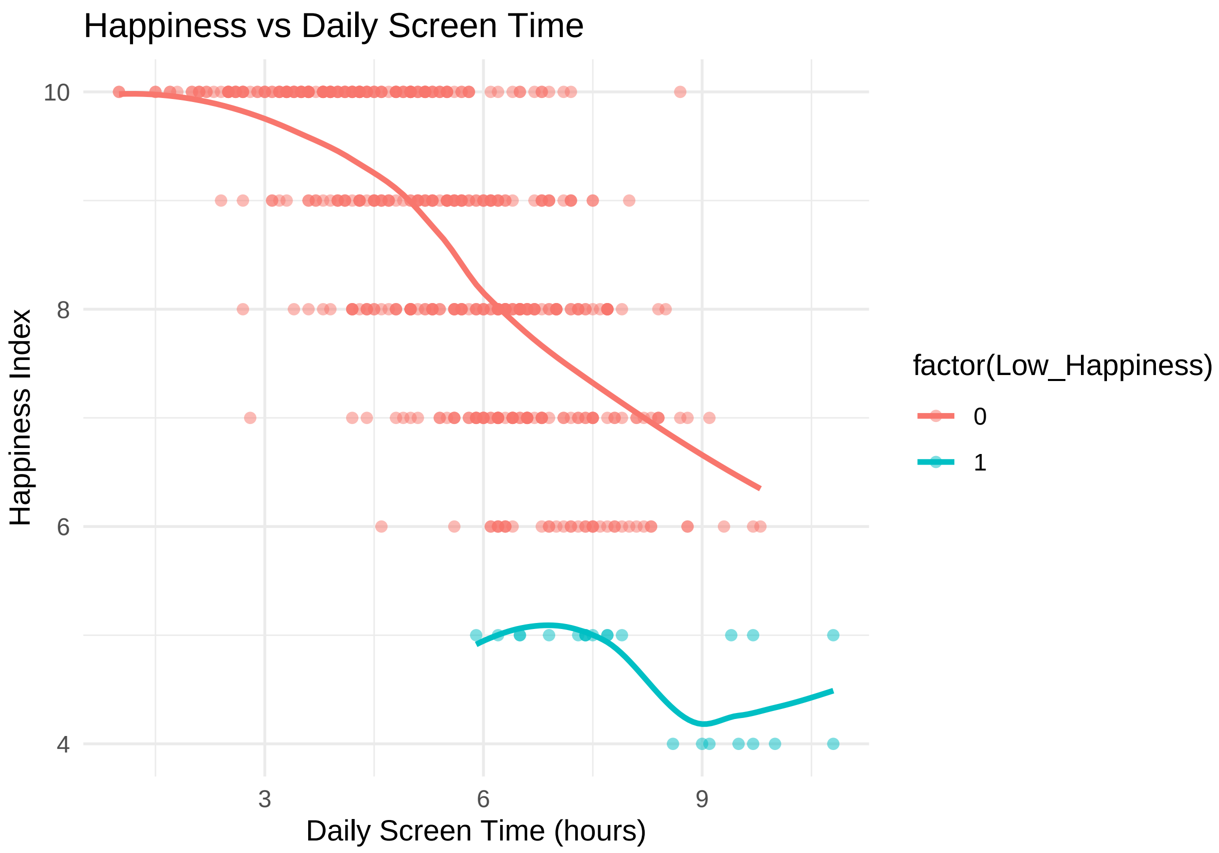
```
##        Metric Value
## 1    Accuracy 0.970
## 2 Sensitivity 0.522
## 3 Specificity 0.992
```

Visualizations

A series of graphics was created to help explain correlations between predictor variables and happiness similar to that enabled by the Shiny application.

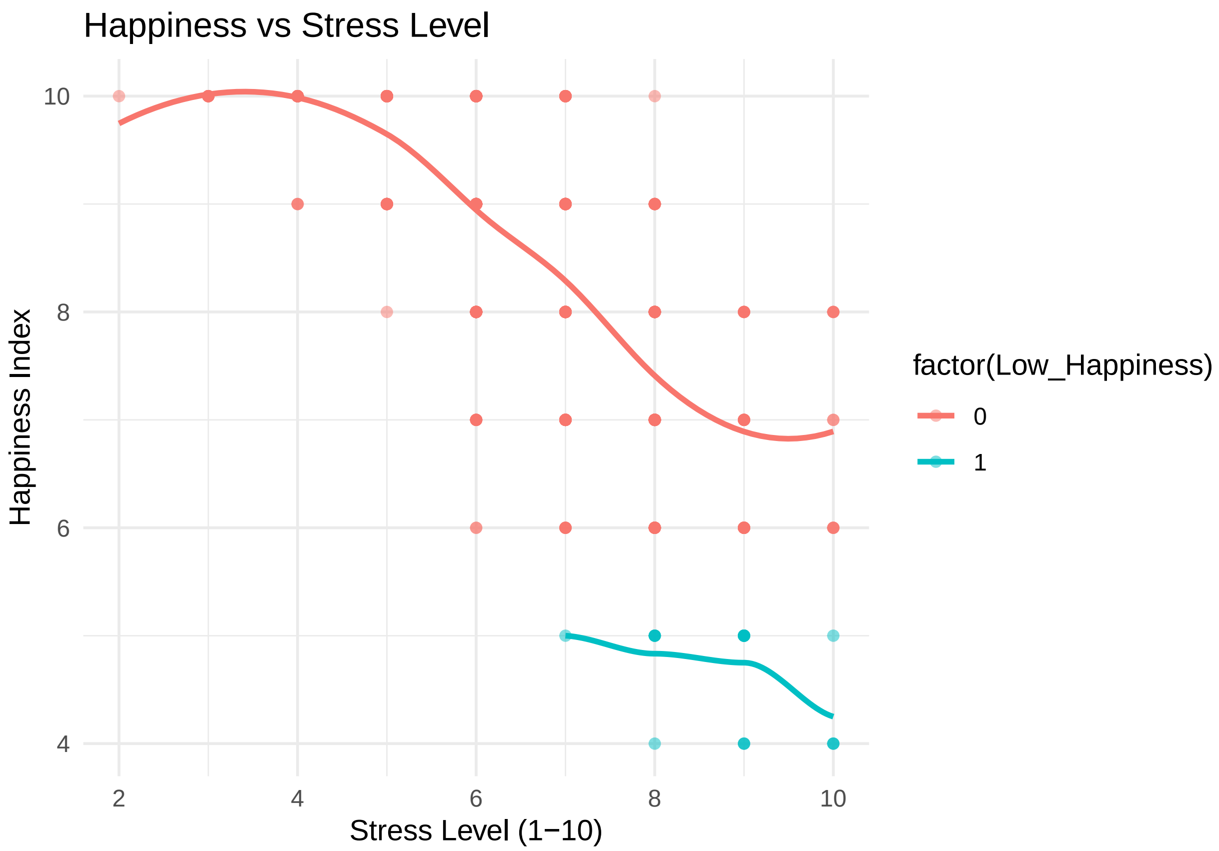
Happiness vs Daily Screen Time

```r
ggplot(df, aes(x = Daily_Screen_Time_hrs, y = Happiness_Index,
               color = factor(Low_Happiness))) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess", se = FALSE) +
  theme_minimal() +
  labs(
    title = "Happiness vs Daily Screen Time",
    x = "Daily Screen Time (hours)",
    y = "Happiness Index"
  )
```



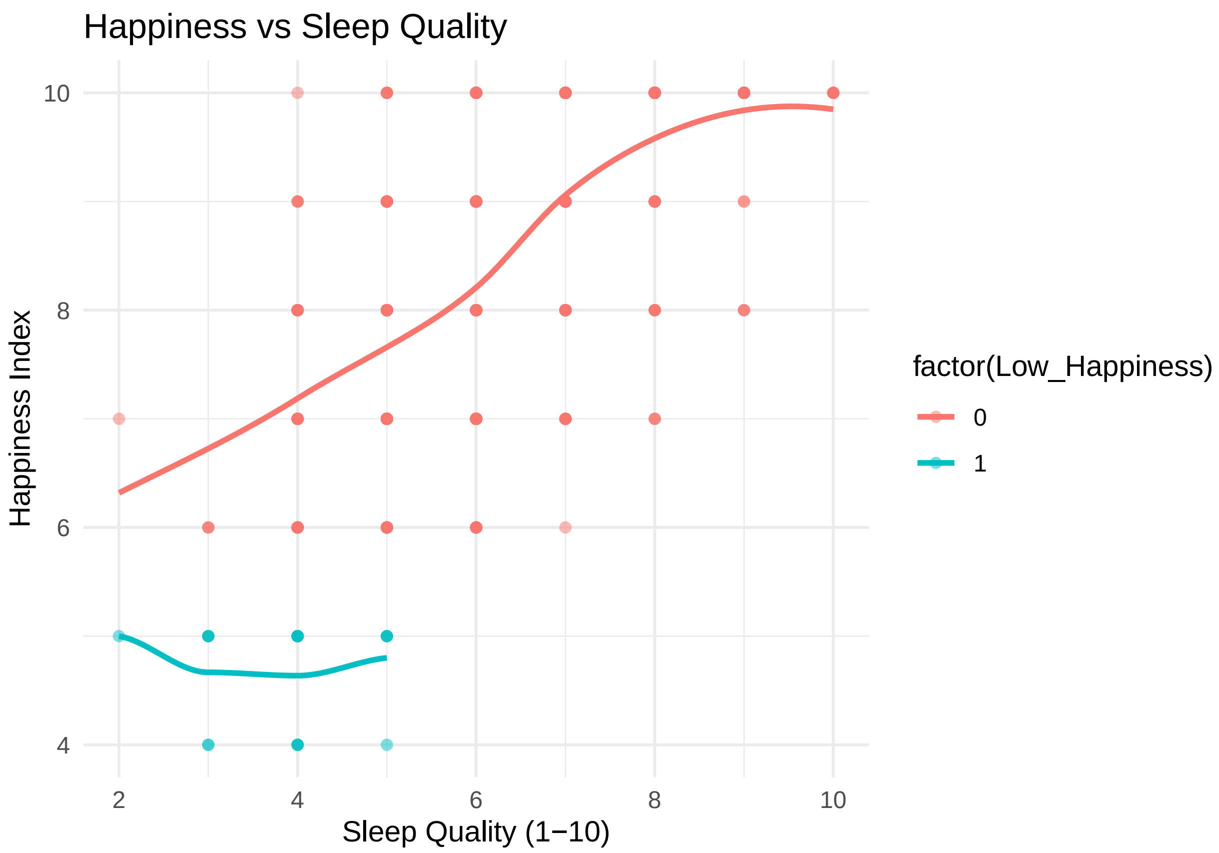
Happiness vs Stress Level

```r
ggplot(df, aes(x = Stress_Level, y = Happiness_Index,
               color = factor(Low_Happiness))) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess", se = FALSE) +
  theme_minimal() +
  labs(
    title = "Happiness vs Stress Level",
    x = "Stress Level (1-10)",
    y = "Happiness Index"
  )
```



Happiness vs Sleep Quality

```r
ggplot(df, aes(x = Sleep_Quality, y = Happiness_Index,
               color = factor(Low_Happiness))) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess", se = FALSE) +
  theme_minimal() +
  labs(
    title = "Happiness vs Sleep Quality",
    x = "Sleep Quality (1-10)",
    y = "Happiness Index"
  )
```



Findings

The overall outcome from the logistic regression and visualization analysis suggests several important associations. Stress Level is found to be the greatest predictor for Low__Happiness, and higher levels of this variable do increase the predictiveness for Low__Happiness. Sleep Quality is found to be remarkably negatively associated with Low__Happiness, which suggests that greater sleep is associated with greater happiness. Exercise Frequency is found to be negatively associated with Low__Happiness and thus supports the known relationship between exercise and happiness. Daily Screen Time is found to be less associated with Low__Happiness, and higher screen time is close to being associated with lower happiness. The classifier performs well on discrimination tasks, as is apparent from the rate of accuracy, specificity, and sensitivity measurements, and this provides evidence on the applicability of the model to this particular dataset. The Shiny application tool allows for interactive analysis and visualization of these results and various parameters.

Conclusion

This project is exemplary of the complete end-to-end process for data science analysis and specifically the application developed and deployed for interactive visualization and analysis with the Shiny application. The analysis that is carried out identifies the significant impact that stress levels, sleep quality, and exercise activities and behaviors have on overall mental well-being and highlights correlations between various online behaviors and happiness levels. The Shiny application is important for offering interactive exploration and analysis between the various significant factors and the impact on `Low_Happiness`.

References

Kaggle Dataset: https://www.kaggle.com/datasets/suchintan/mental-health-social-media-balance-dataset
R Packages Used: dplyr, ggplot2, shiny, stats