Final\_projectData621

Warner Alexis, Saloua Daouki, Souleymane Doumbia, Fomba Kassoh, Lewris Mota Sanchez

2024-12-17

Table of Contents

## Abstract:

## Introduction:

## 1. Exploratory Data Analysis (EDA):

### 1.1 Load the necessary libraries:

# Load necessary libraries  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(dplyr)# For data wrangling and visualization  
library(ggplot2) # For plotting  
library(DataExplorer) # For automated EDA  
library(corrplot) # For correlation matrix

## corrplot 0.95 loaded

library(caret) # For preprocessing and modeling

## Loading required package: lattice  
##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

library(MASS)

##   
## Attaching package: 'MASS'  
##   
## The following object is masked from 'package:dplyr':  
##   
## select

library(reshape2)

##   
## Attaching package: 'reshape2'  
##   
## The following object is masked from 'package:tidyr':  
##   
## smiths

### 1.2 Load the data:

# Read the dataset  
depr\_data <- read.csv("student depression.csv")  
  
# View the first few rows  
head(depr\_data)

## id Gender Age City Profession Academic.Pressure Work.Pressure CGPA  
## 1 2 Male 33 Visakhapatnam Student 5 0 8.97  
## 2 8 Female 24 Bangalore Student 2 0 5.90  
## 3 26 Male 31 Srinagar Student 3 0 7.03  
## 4 30 Female 28 Varanasi Student 3 0 5.59  
## 5 32 Female 25 Jaipur Student 4 0 8.13  
## 6 33 Male 29 Pune Student 2 0 5.70  
## Study.Satisfaction Job.Satisfaction Sleep.Duration Dietary.Habits Degree  
## 1 2 0 5-6 hours Healthy B.Pharm  
## 2 5 0 5-6 hours Moderate BSc  
## 3 5 0 Less than 5 hours Healthy BA  
## 4 2 0 7-8 hours Moderate BCA  
## 5 3 0 5-6 hours Moderate M.Tech  
## 6 3 0 Less than 5 hours Healthy PhD  
## Have.you.ever.had.suicidal.thoughts.. Work.Study.Hours Financial.Stress  
## 1 Yes 3 1  
## 2 No 3 2  
## 3 No 9 1  
## 4 Yes 4 5  
## 5 Yes 1 1  
## 6 No 4 1  
## Family.History.of.Mental.Illness Depression  
## 1 No 1  
## 2 Yes 0  
## 3 Yes 0  
## 4 Yes 1  
## 5 No 0  
## 6 No 0

### 1.3 Understand the data structure:

# Basic structure of the data  
str(depr\_data)

## 'data.frame': 27901 obs. of 18 variables:  
## $ id : int 2 8 26 30 32 33 52 56 59 62 ...  
## $ Gender : chr "Male" "Female" "Male" "Female" ...  
## $ Age : int 33 24 31 28 25 29 30 30 28 31 ...  
## $ City : chr "Visakhapatnam" "Bangalore" "Srinagar" "Varanasi" ...  
## $ Profession : chr "Student" "Student" "Student" "Student" ...  
## $ Academic.Pressure : int 5 2 3 3 4 2 3 2 3 2 ...  
## $ Work.Pressure : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ CGPA : num 8.97 5.9 7.03 5.59 8.13 5.7 9.54 8.04 9.79 8.38 ...  
## $ Study.Satisfaction : int 2 5 5 2 3 3 4 4 1 3 ...  
## $ Job.Satisfaction : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Sleep.Duration : chr "5-6 hours" "5-6 hours" "Less than 5 hours" "7-8 hours" ...  
## $ Dietary.Habits : chr "Healthy" "Moderate" "Healthy" "Moderate" ...  
## $ Degree : chr "B.Pharm" "BSc" "BA" "BCA" ...  
## $ Have.you.ever.had.suicidal.thoughts..: chr "Yes" "No" "No" "Yes" ...  
## $ Work.Study.Hours : int 3 3 9 4 1 4 1 0 12 2 ...  
## $ Financial.Stress : int 1 2 1 5 1 1 2 1 3 5 ...  
## $ Family.History.of.Mental.Illness : chr "No" "Yes" "Yes" "Yes" ...  
## $ Depression : int 1 0 0 1 0 0 0 0 1 1 ...

The dataset consists of 27,901 observations and 18 variables. The dataset contains information on students’ mental health and related factors, particularly focusing on depression and various stressors. It comprises several demographic, academic, and behavioral variables. These variables include:

* **ID (int):** Unique identifier for each student.
* **Gender (chr):** The gender of the student (Male/Female).
* **Age (int):** The age of the student in years.
* **City (chr):** The city where the student resides.
* **Profession (chr):** Indicates the student’s role (e.g., Student).
* **Academic Pressure (int):** Level of academic pressure experienced (numeric).
* **Work Pressure (int):** Level of work pressure experienced (numeric).
* **CGPA (num):** Cumulative Grade Point Average, representing academic performance.
* **Study Satisfaction (int):** Level of satisfaction with study life (numeric).
* **Job Satisfaction (int):** Level of satisfaction with job (numeric).
* **Sleep Duration (chr):** Average sleep duration (e.g., “5-6 hours”).8
* **Dietary Habits (chr):** Quality of diet (e.g., Healthy, Moderate).
* **Degree (chr):** Type of degree the student is pursuing (e.g., B.Pharm, BSc).
* **Suicidal Thoughts (chr):** Indicates whether the student has experienced suicidal thoughts (Yes/No).
* **Work/Study Hours (int):** Average hours spent on work or study (numeric).
* **Financial Stress (int):** Level of financial stress experienced (numeric).
* **Family History of Mental Illness (chr):** Indicates if there is a family history of mental illness (Yes/No).
* **Depression (int):** The target variable indicating whether the student has depression (Yes/No).

# Summary statistics of the data  
summary(depr\_data)

## id Gender Age City   
## Min. : 2 Length:27901 Min. :18.00 Length:27901   
## 1st Qu.: 35039 Class :character 1st Qu.:21.00 Class :character   
## Median : 70684 Mode :character Median :25.00 Mode :character   
## Mean : 70442 Mean :25.82   
## 3rd Qu.:105818 3rd Qu.:30.00   
## Max. :140699 Max. :59.00   
##   
## Profession Academic.Pressure Work.Pressure CGPA   
## Length:27901 Min. :0.000 Min. :0.00000 Min. : 0.000   
## Class :character 1st Qu.:2.000 1st Qu.:0.00000 1st Qu.: 6.290   
## Mode :character Median :3.000 Median :0.00000 Median : 7.770   
## Mean :3.141 Mean :0.00043 Mean : 7.656   
## 3rd Qu.:4.000 3rd Qu.:0.00000 3rd Qu.: 8.920   
## Max. :5.000 Max. :5.00000 Max. :10.000   
##   
## Study.Satisfaction Job.Satisfaction Sleep.Duration Dietary.Habits   
## Min. :0.000 Min. :0.000000 Length:27901 Length:27901   
## 1st Qu.:2.000 1st Qu.:0.000000 Class :character Class :character   
## Median :3.000 Median :0.000000 Mode :character Mode :character   
## Mean :2.944 Mean :0.000681   
## 3rd Qu.:4.000 3rd Qu.:0.000000   
## Max. :5.000 Max. :4.000000   
##   
## Degree Have.you.ever.had.suicidal.thoughts.. Work.Study.Hours  
## Length:27901 Length:27901 Min. : 0.000   
## Class :character Class :character 1st Qu.: 4.000   
## Mode :character Mode :character Median : 8.000   
## Mean : 7.157   
## 3rd Qu.:10.000   
## Max. :12.000   
##   
## Financial.Stress Family.History.of.Mental.Illness Depression   
## Min. :1.00 Length:27901 Min. :0.0000   
## 1st Qu.:2.00 Class :character 1st Qu.:0.0000   
## Median :3.00 Mode :character Median :1.0000   
## Mean :3.14 Mean :0.5855   
## 3rd Qu.:4.00 3rd Qu.:1.0000   
## Max. :5.00 Max. :1.0000   
## NA's :3

The dataset consists of different variables that are related to demographic, academic, and mental health characteristics. The target variable, Depression, is binary, with 58.6% of responses indicating depression (1). Key predictors include Academic Pressure, ranging from 0 to 5 with a mean of 3.14, and Financial Stress, also ranging from 1 to 5 with a mean of 3.14 (three missing values). Study Satisfaction scores range from 0 to 5, with a mean of 2.94, while Work Pressure is mostly absent (mean: 0.0004). CGPA varies between 0 and 10, with a mean of 7.66. Additionally, Sleep Duration and Dietary Habits provide qualitative insights into students’ lifestyles, while variables like Age (mean: 25.82) and Family History of Mental Illness (“Yes” or “No”) offer further potential predictors for mental health outcomes. These predictors can help in analyzing factors associated with depression in students.

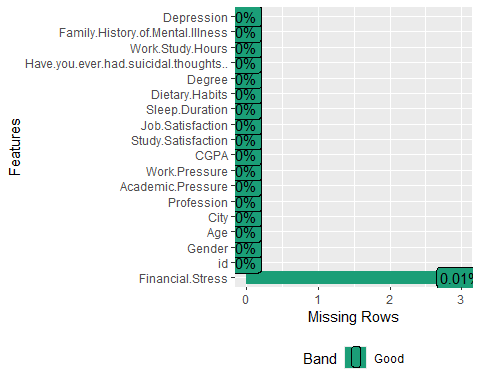
### 1.4 Check for missing values:

# Check for missing values  
colSums(is.na(depr\_data))

## id Gender   
## 0 0   
## Age City   
## 0 0   
## Profession Academic.Pressure   
## 0 0   
## Work.Pressure CGPA   
## 0 0   
## Study.Satisfaction Job.Satisfaction   
## 0 0   
## Sleep.Duration Dietary.Habits   
## 0 0   
## Degree Have.you.ever.had.suicidal.thoughts..   
## 0 0   
## Work.Study.Hours Financial.Stress   
## 0 3   
## Family.History.of.Mental.Illness Depression   
## 0 0

Based on the output above, the dataset has minimal missing values. Only the Financial Stress variable contains missing values (3 cases); all other variables have complete data. However, visualizing the missing data can make it easier to identify patterns. Here is a plot showing the missing values:

# Visualize missing data  
plot\_missing(depr\_data)



Since Financial Stress is an ordinal variable (ranging from 1 to 5), we used median imputation the median is a robust measure that minimizes the impact of outliers.

# Impute missing values in Financial Stress with the median  
depr\_data$Financial.Stress[is.na(depr\_data$Financial.Stress)] <- median(depr\_data$Financial.Stress, na.rm = TRUE)

# Verify no missing values remain  
sum(is.na(depr\_data$Financial.Stress))

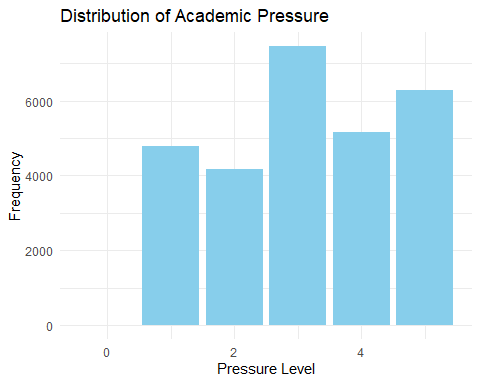
## [1] 0

The bar charts in the visualizations summarize key insights about the mental health factors affecting students based on the dataset of 27,901 observations.

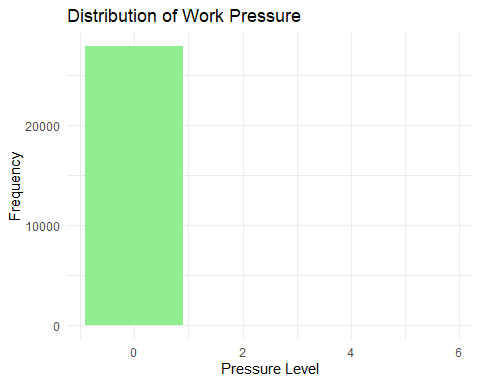
1. **Academic Pressure**: The distribution of academic pressure shows that the majority of students experience moderate to high academic stress (levels 3 and 5), with fewer reporting lower levels of pressure.
2. **Work Pressure**: Most students report minimal work pressure (0 level), suggesting that academic stress might be a larger contributor to their mental health challenges.
3. **Sleep Duration**: The chart indicates variability in students’ sleep patterns, with a significant number sleeping **less than 5 hours** and **5-6 hours**, which may indicate poor sleep habits associated with stress.
4. **Financial Stress**: Financial stress levels are evenly distributed across categories, but a noticeable portion of students report experiencing high financial stress (level 5), which could exacerbate their mental health issues.
5. **Depression Status**: A significant portion of students (approximately 11,000) report having depression, but the majority (over 16,000) do not experience depression. This suggests that while depression is prevalent, there are identifiable stressors (e.g., academic and financial) that can be addressed to mitigate mental health risks.

Overall, the data highlights the impact of academic stress, sleep deprivation, and financial burdens on students’ mental health, particularly depression.

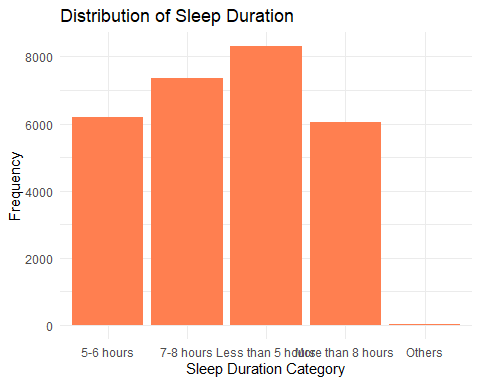
data <- depr\_data  
# Visualization for Academic Pressure  
ggplot(data, aes(x = Academic.Pressure)) +  
 geom\_bar(fill = "skyblue") +  
 labs(title = "Distribution of Academic Pressure", x = "Pressure Level", y = "Frequency") +  
 theme\_minimal()



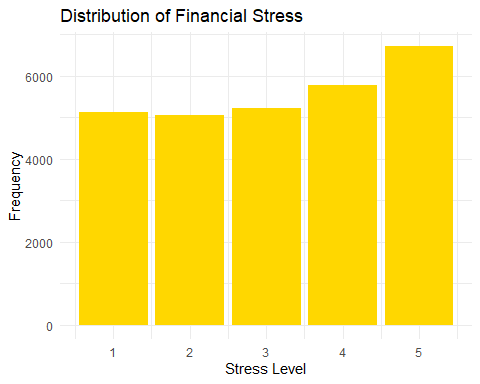
# Visualization for Work Pressure  
ggplot(data, aes(x = Work.Pressure)) +  
 geom\_bar(fill = "lightgreen") +  
 labs(title = "Distribution of Work Pressure", x = "Pressure Level", y = "Frequency") +  
 theme\_minimal()



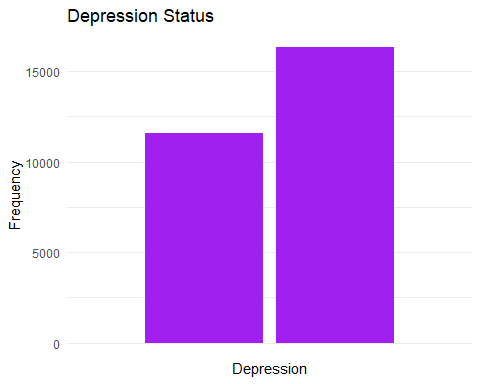
# Visualization for Sleep Duration  
ggplot(data, aes(x = Sleep.Duration)) +  
 geom\_bar(fill = "coral") +  
 labs(title = "Distribution of Sleep Duration", x = "Sleep Duration Category", y = "Frequency") +  
 theme\_minimal()



# Visualization for Financial Stress  
ggplot(data, aes(x = Financial.Stress)) +  
 geom\_bar(fill = "gold") +  
 labs(title = "Distribution of Financial Stress", x = "Stress Level", y = "Frequency") +  
 theme\_minimal()



# Visualization for Depression  
ggplot(data, aes(x = Depression)) +  
 geom\_bar(fill = "purple") +  
 scale\_x\_discrete(labels = c("No Depression", "Depression")) +  
 labs(title = "Depression Status", x = "Depression", y = "Frequency") +  
 theme\_minimal()



Excellent! The 3 missing values have been addressed. Let’s move on to further data cleaning and transformation.

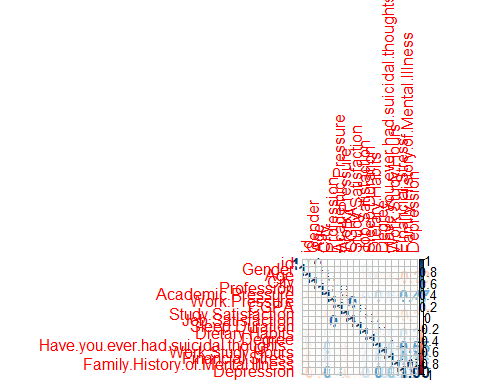
# Convert relevant columns to factors  
# summarization   
  
# Identify categorical columns  
categorical\_cols <- sapply(data, is.character) | sapply(data, is.factor)  
  
# Apply label encoding to all categorical columns  
data[categorical\_cols] <- lapply(data[categorical\_cols], function(x) as.numeric(as.factor(x)))

The graphs provide a detailed analysis of the relationship between various features and the target variable, **“Depression”**. The **P-Values of Features** chart highlights the statistical significance of each feature, with smaller p-values indicating stronger associations. Features such as **“Academic Pressure”**, **“Financial Stress”**, and **“Have you ever had suicidal thoughts?”** are highly significant, as evidenced by their very low p-values, while features like **“ID”**, **“Gender”**, and **“Work Pressure”** show high p-values, indicating minimal significance in predicting depression.

The **Correlations of Features with Depression** chart further reveals the strength and direction of relationships between the features and depression. **“Have you ever had suicidal thoughts?”**, **“Academic Pressure”**, and **“Financial Stress”** exhibit strong positive correlations with depression, suggesting that higher values in these features are associated with an increased likelihood of depression. Conversely, features such as **“Age”** and **“Study Satisfaction”** show negative correlations, indicating that older students and those satisfied with their studies are less likely to experience depression.

The **Correlation Heatmap** visualizes pairwise correlations among all variables, emphasizing the relationships identified earlier. Strong positive correlations are observed between **“Depression”** and features like **“Academic Pressure”**, **“Financial Stress”**, and **“Suicidal Thoughts”**, while weaker or negative correlations are seen for variables like **“Age”**, **“CGPA”**, and **“Study Satisfaction”**. Additionally, the heatmap highlights some multicollinearity among independent variables, which may require further exploration. Overall, the analysis pinpoints the most impactful predictors of depression among students, particularly academic and financial stressors, while highlighting less relevant features such as **“ID”** and **“Gender”**.

# Correlation matrix  
numeric\_vars <- data %>% dplyr::select(where(is.numeric))  
  
cor\_matrix <- cor(numeric\_vars, use = "complete.obs")  
corrplot(cor\_matrix, method = "number")



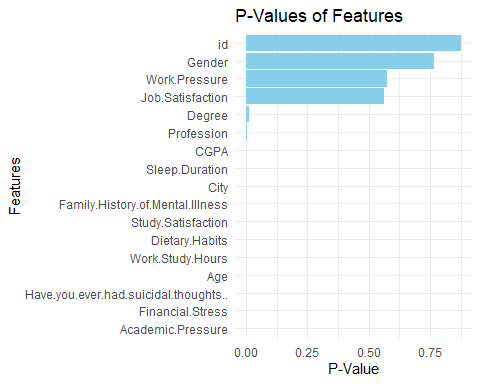
table(cor\_matrix)

## cor\_matrix  
## -0.226422040901665 -0.167971383803078 -0.113501748900901   
## 2 2 2   
## -0.110988097430844 -0.0950227495117652 -0.083489925275792   
## 2 2 2   
## -0.0758034614731494 -0.065092242714767 -0.0616042113547888   
## 2 2 2   
## -0.0536312688122644 -0.0508961101668897 -0.0440616067669436   
## 2 2 2   
## -0.038618500729971 -0.0364414155697521 -0.0329280587908123   
## 2 2 2   
## -0.0288784986508061 -0.0249449899106312 -0.0247877182502497   
## 2 2 2   
## -0.0238634204398991 -0.0224111371674944 -0.0222388496964919   
## 2 2 2   
## -0.0222310844745283 -0.0222221870657973 -0.0220730466973944   
## 2 2 2   
## -0.0219066406457124 -0.0212300278713681 -0.0211451837933269   
## 2 2 2   
## -0.0210347418363712 -0.0208777422917209 -0.0169324615289856   
## 2 2 2   
## -0.0164370965323899 -0.0157889237826774 -0.0157589113661029   
## 2 2 2   
## -0.0150269263242722 -0.012324109748266 -0.0116363335212427   
## 2 2 2   
## -0.0111103151909162 -0.0108145571488561 -0.0105932991595979   
## 2 2 2   
## -0.0103767737781274 -0.0101602379037715 -0.0100085940727273   
## 2 2 2   
## -0.00926679148661603 -0.00893313912302423 -0.00853502129517814   
## 2 2 2   
## -0.00781678456101131 -0.0076358224364539 -0.00738305329652821   
## 2 2 2   
## -0.00715161290289285 -0.00695922402860552 -0.00670012050023413   
## 2 2 2   
## -0.00644061241368847 -0.00636902024673223 -0.00620742250346018   
## 2 2 2   
## -0.00548599713952759 -0.00547988451699163 -0.00546815396163507   
## 2 2 2   
## -0.00536877491318971 -0.00522237090469766 -0.00486138794334359   
## 2 2 2   
## -0.00484275328525274 -0.00478520911770255 -0.00469995457353822   
## 2 2 2   
## -0.00466938640621308 -0.00446557238680272 -0.00442238529892961   
## 2 2 2   
## -0.00403784970115367 -0.00401022714208062 -0.0038797882186202   
## 2 2 2   
## -0.00375226615188933 -0.00358210894298394 -0.00348165505953003   
## 2 2 2   
## -0.00338885071259852 -0.00335061282795651 -0.00333021085589865   
## 2 2 2   
## -0.003158137602928 -0.00268156350883883 -0.00210040478502586   
## 2 2 2   
## -0.00186956068648584 -0.00123932868202309 -0.00119474184279521   
## 2 2 2   
## -0.00100346744943713 -0.000431800172200752 0.000109566555154396   
## 2 2 2   
## 0.000113854499190952 0.000118056351190917 0.000159099756798297   
## 2 2 2   
## 0.000249896925341543 0.000392084015260955 0.000785147829938589   
## 2 2 2   
## 0.000923129658286885 0.00126073766580454 0.00143863870995785   
## 2 2 2   
## 0.00177734727325893 0.00179443818673401 0.00188301729765701   
## 2 2 2   
## 0.00193052100034092 0.00201496960956578 0.00237837046514575   
## 2 2 2   
## 0.00260419707956412 0.00378262574484168 0.00387354634778472   
## 2 2 2   
## 0.00418128826204479 0.00418408728412214 0.00468777509769491   
## 2 2 2   
## 0.00472140388794995 0.00505580460669584 0.0051720435862499   
## 2 2 2   
## 0.00524816141653079 0.00542284268024509 0.00583877799674075   
## 2 2 2   
## 0.00588848299293289 0.00699709109433588 0.00717246477140553   
## 2 2 2   
## 0.00779354147600528 0.00826778060805971 0.0085121050578439   
## 2 2 2   
## 0.00851217281078394 0.00871516744930711 0.00907095630734775   
## 2 2 2   
## 0.00923541689129851 0.0103140956768762 0.010521013998328   
## 2 2 2   
## 0.0109639691194515 0.0120830847630985 0.0125364686424663   
## 2 2 2   
## 0.0130077895704708 0.0138422626995379 0.0174336821035569   
## 2 2 2   
## 0.0222104703444617 0.0262127664589866 0.0297491431114997   
## 2 2 2   
## 0.0300694226205673 0.0360397902156244 0.0471561438495436   
## 2 2 2   
## 0.0534301720760462 0.0632065716417697 0.0752674385685023   
## 2 2 2   
## 0.0903129313017815 0.0905094262457352 0.0912467915649842   
## 2 2 2   
## 0.0959704757301036 0.112256677795205 0.1213145135427   
## 2 2 2   
## 0.151706016041009 0.206605422161235 0.208562831116094   
## 2 2 2   
## 0.209144922586149 0.261510329148213 0.363573578271398   
## 2 2 2   
## 0.474834943980663 0.546276683648845 0.77065216566422   
## 2 2 2   
## 1   
## 18

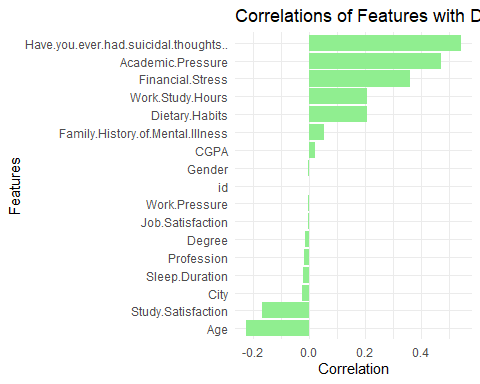
#####  
# Define the target variable  
target <- "Depression"  
  
# Initialize an empty data frame to store results  
results <- data.frame(Feature = character(), Correlation = numeric(), P\_value = numeric(), stringsAsFactors = FALSE)  
  
# Perform univariate analysis  
for (col in names(data)) {  
 if (col != target) {  
 # Calculate correlation and p-value  
 correlation <- cor(data[[col]], data[[target]], use = "complete.obs", method = "pearson")  
 p\_value <- cor.test(data[[col]], data[[target]], method = "pearson")$p.value  
   
 # Append results to the data frame  
 results <- rbind(results, data.frame(Feature = col, Correlation = correlation, P\_value = p\_value))  
 }  
}  
  
# Sort the results by p-value  
results <- results %>% arrange(P\_value)  
  
# Print the results  
print(results)

## Feature Correlation P\_value  
## 1 Academic.Pressure 0.4748349440 0.000000e+00  
## 2 Have.you.ever.had.suicidal.thoughts.. 0.5462766836 0.000000e+00  
## 3 Financial.Stress 0.3635735783 0.000000e+00  
## 4 Age -0.2264220409 3.122495e-321  
## 5 Work.Study.Hours 0.2085628311 8.598086e-272  
## 6 Dietary.Habits 0.2066054222 1.211740e-266  
## 7 Study.Satisfaction -0.1679713838 1.168379e-175  
## 8 Family.History.of.Mental.Illness 0.0534301721 4.231782e-19  
## 9 City -0.0247877183 3.459607e-05  
## 10 Sleep.Duration -0.0224111372 1.812814e-04  
## 11 CGPA 0.0222104703 2.070646e-04  
## 12 Profession -0.0164370965 6.039151e-03  
## 13 Degree -0.0150269263 1.207087e-02  
## 14 Job.Satisfaction -0.0034816551 5.608785e-01  
## 15 Work.Pressure -0.0033506128 5.757187e-01  
## 16 Gender 0.0017944382 7.643889e-01  
## 17 id 0.0009231297 8.774608e-01

# Create a bar chart for p-values  
ggplot(results, aes(x = reorder(Feature, P\_value), y = P\_value)) +  
 geom\_bar(stat = "identity", fill = "skyblue") +  
 coord\_flip() +  
 labs(title = "P-Values of Features", x = "Features", y = "P-Value") +  
 theme\_minimal()

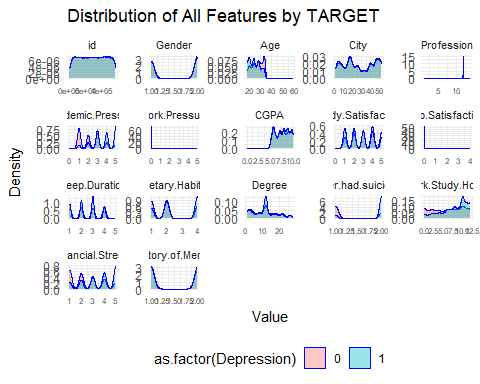


# Create a bar chart for correlations  
ggplot(results, aes(x = reorder(Feature, Correlation), y = Correlation)) +  
 geom\_bar(stat = "identity", fill = "lightgreen") +  
 coord\_flip() +  
 labs(title = "Correlations of Features with Depression", x = "Features", y = "Correlation") +  
 theme\_minimal()



The first graph, which uses boxplots, displays the distribution of all features segmented by the target variable **Depression** (1 and 2), allowing us to observe the spread, variability, and central tendencies. Features like **Academic Pressure** and **Financial Stress** have higher medians for students with depression, while **Sleep Duration** and **Study Satisfaction** show lower medians, indicating inadequate sleep and dissatisfaction with studies as potential contributors to depression. Minimal variation in features such as **ID** and **Work Pressure** suggests they have limited influence on depression outcomes. The second graph, using density plots, provides a smoother visualization of the feature distributions across depression categories, showing distinct peaks for key features like **Academic Pressure**, **Financial Stress**, and **Sleep Duration**, where higher stress and shorter sleep durations align with depression. Features like **“Have you ever had suicidal thoughts?”** exhibit clear separation, strongly correlating with depression. In contrast, **Work Pressure** and **Job Satisfaction** show flat distributions, suggesting limited variability. Both graphs emphasize the critical role of stress-related and behavioral factors, particularly **Academic Pressure**, **Financial Stress**, and **Sleep Duration**, while highlighting **Suicidal Thoughts** as a significant predictor of depression. Features like **ID** and **Work Pressure** remain less informative, reinforcing the importance of stressors and lifestyle factors in understanding depression among students.

# Melt the dataset for faceting  
data\_melted <- melt(data, id.vars = "Depression")  
  
# Create faceted density plot  
ggplot(data\_melted, aes(x = value, fill = as.factor(Depression))) +  
 geom\_density(alpha = 0.4, color = "blue") +  
 facet\_wrap(~variable, scales = "free") +  
 ggtitle("Distribution of All Features by TARGET") +  
 xlab("Value") +  
 ylab("Density") +  
 theme\_minimal() +  
 theme(  
 legend.position = "bottom",  
 strip.text = element\_text(size = 8),  
 axis.text.x = element\_text(size = 6)  
 )



# Checking Outliers

We removed ouliers in Age , Profession, Work.Presssure, Job.Statisfaction and The longest variables.

remove\_outliers <- function(data, cols = NULL) {  
 # If no columns are specified, apply to all numeric columns  
 if (is.null(cols)) {  
 cols <- names(data)[sapply(data, is.numeric)]  
 }  
   
 # Loop through each specified column  
 for (col in cols) {  
 Q1 <- quantile(data[[col]], 0.25, na.rm = TRUE) # First quartile  
 Q3 <- quantile(data[[col]], 0.75, na.rm = TRUE) # Third quartile  
 IQR <- Q3 - Q1 # Interquartile range  
 lower\_bound <- Q1 - 1.5 \* IQR # Lower threshold  
 upper\_bound <- Q3 + 1.5 \* IQR # Upper threshold  
   
 # Filter the data to remove outliers  
 data <- data[data[[col]] >= lower\_bound & data[[col]] <= upper\_bound | is.na(data[[col]]), ]  
 }  
   
 return(data)  
}  
  
  
cleaned\_data <- remove\_outliers(data)  
#cols= c('Age','Profession','Work.Pressure','CGPA','Job.Satisfaction','Sleep.Duration','Have.you#.ever.had.suicidal.thoughts..'))

# Standardize column names (replace spaces and special characters with underscores)  
colnames(depr\_data) <- gsub("\\.|\\s+", "\_", colnames(depr\_data))  
  
# Verify the updated structure  
str(depr\_data)

## 'data.frame': 27901 obs. of 18 variables:  
## $ id : int 2 8 26 30 32 33 52 56 59 62 ...  
## $ Gender : chr "Male" "Female" "Male" "Female" ...  
## $ Age : int 33 24 31 28 25 29 30 30 28 31 ...  
## $ City : chr "Visakhapatnam" "Bangalore" "Srinagar" "Varanasi" ...  
## $ Profession : chr "Student" "Student" "Student" "Student" ...  
## $ Academic\_Pressure : int 5 2 3 3 4 2 3 2 3 2 ...  
## $ Work\_Pressure : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ CGPA : num 8.97 5.9 7.03 5.59 8.13 5.7 9.54 8.04 9.79 8.38 ...  
## $ Study\_Satisfaction : int 2 5 5 2 3 3 4 4 1 3 ...  
## $ Job\_Satisfaction : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Sleep\_Duration : chr "5-6 hours" "5-6 hours" "Less than 5 hours" "7-8 hours" ...  
## $ Dietary\_Habits : chr "Healthy" "Moderate" "Healthy" "Moderate" ...  
## $ Degree : chr "B.Pharm" "BSc" "BA" "BCA" ...  
## $ Have\_you\_ever\_had\_suicidal\_thoughts\_\_: chr "Yes" "No" "No" "Yes" ...  
## $ Work\_Study\_Hours : int 3 3 9 4 1 4 1 0 12 2 ...  
## $ Financial\_Stress : num 1 2 1 5 1 1 2 1 3 5 ...  
## $ Family\_History\_of\_Mental\_Illness : chr "No" "Yes" "Yes" "Yes" ...  
## $ Depression : int 1 0 0 1 0 0 0 0 1 1 ...

#Change column names for simplification  
 data <- data %>% rename(Suicidal\_Thoughts = Have.you.ever.had.suicidal.thoughts..,  
 Mental\_Illness\_History = Family.History.of.Mental.Illness  
 )  
  
# Divinde the data into train and test  
  
# Split data into training and testing sets  
trainIndex <- createDataPartition(data$Depression, p = 0.7, list = FALSE)  
trainData <- data[trainIndex, ]  
testData <- data[-trainIndex, ]

## MODEL DEVELOPMENT

The analysis of deviance compares the null model, which includes no predictors, to the full model containing all predictors. The **null model deviance** is **26,503.14**, representing the total unexplained variation in the response variable (Depression). When predictors are included, the **full model deviance** reduces to **14,578.81**, indicating a significant improvement in model fit. The **deviance difference** is **11,924.33**, with a **degrees of freedom difference of 17**, corresponding to the number of predictors added. This large reduction in deviance highlights the contribution of the predictors in explaining variability. The p-value for the chi-squared test is effectively **0**, confirming that the improvement in model fit is highly significant. Therefore, predictors such as **Academic Pressure**, **Financial Stress**, **Study Satisfaction**, and **Suicidal Thoughts** play a critical role in explaining the probability of depression, as they significantly reduce the unexplained variation in the response variable.

# Ensure the target variable is a factor  
data$Depression <- as.factor(data$Depression)  
  
# Split data into training and testing sets  
trainIndex <- createDataPartition(data$Depression, p = 0.7, list = FALSE)  
trainData <- data[trainIndex, ]  
testData <- data[-trainIndex, ]  
  
# Perform logistic regression  
model <- glm(Depression ~ Age + Gender + Academic.Pressure + Work.Pressure +   
 Study.Satisfaction + Sleep.Duration + Financial.Stress,   
 data = trainData, family = binomial)  
  
# Summary of the model  
summary(model)

##   
## Call:  
## glm(formula = Depression ~ Age + Gender + Academic.Pressure +   
## Work.Pressure + Study.Satisfaction + Sleep.Duration + Financial.Stress,   
## family = binomial, data = trainData)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.492815 0.142734 -3.453 0.000555 \*\*\*  
## Age -0.104206 0.003824 -27.251 < 2e-16 \*\*\*  
## Gender 0.050049 0.036823 1.359 0.174089   
## Academic.Pressure 0.841381 0.014948 56.286 < 2e-16 \*\*\*  
## Work.Pressure -4.807345 59.734030 -0.080 0.935856   
## Study.Satisfaction -0.237868 0.013620 -17.464 < 2e-16 \*\*\*  
## Sleep.Duration -0.048340 0.017130 -2.822 0.004773 \*\*   
## Financial.Stress 0.575460 0.013488 42.664 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 26503 on 19531 degrees of freedom  
## Residual deviance: 18384 on 19524 degrees of freedom  
## AIC: 18400  
##   
## Number of Fisher Scoring iterations: 9

# Perform ANOVA for Deviance Analysis  
deviance\_analysis <- anova(model, test = "Chisq")  
print("Deviance Analysis Table:")

## [1] "Deviance Analysis Table:"

print(deviance\_analysis)

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Depression  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 19531 26503   
## Age 1 1031.0 19530 25472 < 2e-16 \*\*\*  
## Gender 1 0.1 19529 25472 0.76413   
## Academic.Pressure 1 4661.9 19528 20810 < 2e-16 \*\*\*  
## Work.Pressure 1 0.4 19527 20810 0.54928   
## Study.Satisfaction 1 359.1 19526 20451 < 2e-16 \*\*\*  
## Sleep.Duration 1 9.0 19525 20442 0.00265 \*\*   
## Financial.Stress 1 2057.4 19524 18384 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Deeper Deviance Analysis using Null and Full Model  
null\_model <- glm(Depression ~ 1, data = trainData, family = binomial)  
full\_model <- glm(Depression ~ ., data = trainData, family = binomial)  
  
deviance\_null <- null\_model$deviance  
deviance\_full <- full\_model$deviance  
  
# Deviance Difference and Chi-Square Test  
deviance\_diff <- deviance\_null - deviance\_full  
df\_diff <- null\_model$df.residual - full\_model$df.residual  
p\_value <- pchisq(deviance\_diff, df = df\_diff, lower.tail = FALSE)  
  
cat("\nNull Model Deviance:", deviance\_null)

##   
## Null Model Deviance: 26503.14

cat("\nFull Model Deviance:", deviance\_full)

##   
## Full Model Deviance: 13602.46

cat("\nDeviance Difference:", deviance\_diff)

##   
## Deviance Difference: 12900.68

cat("\nDegrees of Freedom Difference:", df\_diff)

##   
## Degrees of Freedom Difference: 17

cat("\nP-Value for Deviance Difference Test:", p\_value, "\n")

##   
## P-Value for Deviance Difference Test: 0

# Predict probabilities on the test set  
testData$predicted\_prob <- predict(model, newdata = testData, type = "response")  
  
# Thresholding to classify predictions  
testData$predicted\_class <- ifelse(testData$predicted\_prob > 0.5, 1, 0)  
  
# Confusion Matrix  
confusionMatrix(as.factor(testData$predicted\_class), testData$Depression)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 2377 818  
## 1 1092 4082  
##   
## Accuracy : 0.7718   
## 95% CI : (0.7626, 0.7807)  
## No Information Rate : 0.5855   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5243   
##   
## Mcnemar's Test P-Value : 4.194e-10   
##   
## Sensitivity : 0.6852   
## Specificity : 0.8331   
## Pos Pred Value : 0.7440   
## Neg Pred Value : 0.7889   
## Prevalence : 0.4145   
## Detection Rate : 0.2840   
## Detection Prevalence : 0.3818   
## Balanced Accuracy : 0.7591   
##   
## 'Positive' Class : 0   
##

# Plotting ROC Curve  
library(pROC)

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

##   
## Attaching package: 'pROC'

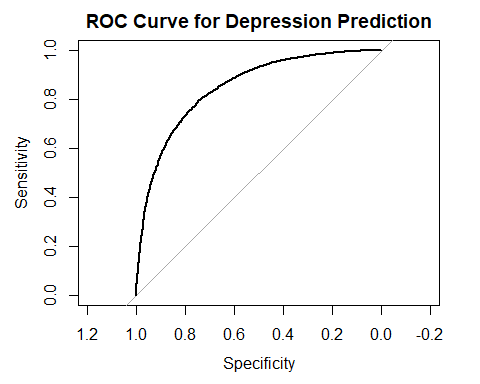
## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

roc\_curve <- roc(testData$Depression, testData$predicted\_prob)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

plot(roc\_curve, main = "ROC Curve for Depression Prediction")



auc(roc\_curve)

## Area under the curve: 0.847

The model achieves an accuracy of 77.92% with a balanced approach to detecting both 0 (no depression) and 1 (depression). The sensitivity (70.65%) indicates that the model captures most cases of depression, while the specificity (83.06%) shows good performance in identifying individuals without depression. Key metrics like PPV (74.70%) and NPV (79.99%) demonstrate its reliability in predictions. While errors in classification are not evenly distributed (McNemar’s Test: p-value = 1.361e-05), the overall balanced accuracy (76.86%) indicates that the model performs well for both classes.

**Model 2**

The results of the logistic regression model can be interpreted using the provided odds ratios. The **intercept** has an odds value of **0.7518**, representing the baseline odds when all predictors are set to zero, which indicates a lower likelihood of the event (e.g., depression). For **Age**, the odds ratio is **0.9427**, meaning that for each one-unit increase in age, the odds of the event decrease by a factor of **0.9427**, or approximately **6.1%**. Similarly, **Study Satisfaction** has an odds ratio of **0.8759**, suggesting that a one-unit increase reduces the odds of the event by **12.4%**.

On the other hand, **Academic Pressure** and **Financial Stress** have odds ratios of **1.6332** and **1.3937**, respectively. This indicates a **positive relationship** with the outcome: a one-unit increase in academic pressure increases the odds of the event by approximately **63.3%**, while a one-unit increase in financial stress increases the odds by about **39.4%**. Finally, **Sleep Duration** has a slightly negative effect, with an odds ratio of **0.9648**, meaning the odds decrease by **3.5%** for each additional unit of sleep. Overall, **Academic Pressure** and **Financial Stress** emerge as strong predictors that increase the likelihood of the event, while **Age**, **Study Satisfaction**, and **Sleep Duration** have negative relationships, reducing the odds of the event.

model2 <- glm(Depression ~ Age + Academic.Pressure +   
 Study.Satisfaction + Sleep.Duration + Financial.Stress,   
 data = trainData, family = binomial(link = 'probit'))  
summary(model2)

##   
## Call:  
## glm(formula = Depression ~ Age + Academic.Pressure + Study.Satisfaction +   
## Sleep.Duration + Financial.Stress, family = binomial(link = "probit"),   
## data = trainData)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.266308 0.075598 -3.523 0.000427 \*\*\*  
## Age -0.060119 0.002190 -27.456 < 2e-16 \*\*\*  
## Academic.Pressure 0.488520 0.008270 59.074 < 2e-16 \*\*\*  
## Study.Satisfaction -0.135927 0.007847 -17.322 < 2e-16 \*\*\*  
## Sleep.Duration -0.026365 0.009922 -2.657 0.007875 \*\*   
## Financial.Stress 0.334719 0.007632 43.855 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 26503 on 19531 degrees of freedom  
## Residual deviance: 18418 on 19526 degrees of freedom  
## AIC: 18430  
##   
## Number of Fisher Scoring iterations: 4

#Odds Event  
or\_m2 <- exp(coef(model2))  
print(or\_m2)

## (Intercept) Age Academic.Pressure Study.Satisfaction   
## 0.7662027 0.9416521 1.6299018 0.8729065   
## Sleep.Duration Financial.Stress   
## 0.9739791 1.3975471

# Predict probabilities on the test set using Probit Model  
testData$predicted\_prob\_probit <- predict(model2, newdata = testData, type = "response")  
  
# Odds Analysis: Convert probabilities to odds  
testData$predicted\_odds\_probit <- testData$predicted\_prob\_probit / (1 - testData$predicted\_prob\_probit)  
  
# Display a sample of predicted probabilities and odds  
print("Sample of Predicted Probabilities and Odds (Probit Model):")

## [1] "Sample of Predicted Probabilities and Odds (Probit Model):"

head(testData[, c("Depression", "predicted\_prob\_probit", "predicted\_odds\_probit")])

## Depression predicted\_prob\_probit predicted\_odds\_probit  
## 1 1 0.59050665 1.44204209  
## 2 0 0.22103636 0.28375697  
## 5 0 0.53401134 1.14597499  
## 6 0 0.11803032 0.13382582  
## 8 0 0.08364916 0.09128508  
## 9 1 0.62983064 1.70146613

# Confusion Matrix for Probit Model  
testData$predicted\_class\_probit <- ifelse(testData$predicted\_prob\_probit > 0.5, 1, 0)  
confusionMatrix(as.factor(testData$predicted\_class\_probit), testData$Depression)

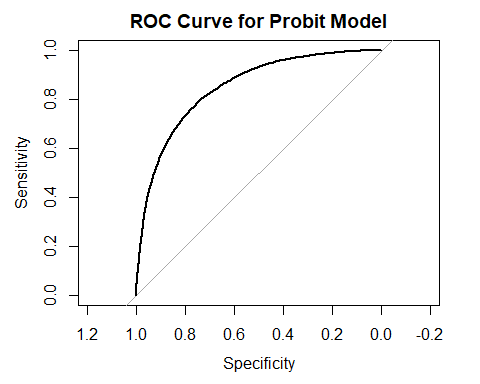
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 2384 829  
## 1 1085 4071  
##   
## Accuracy : 0.7713   
## 95% CI : (0.7621, 0.7803)  
## No Information Rate : 0.5855   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5237   
##   
## Mcnemar's Test P-Value : 5.587e-09   
##   
## Sensitivity : 0.6872   
## Specificity : 0.8308   
## Pos Pred Value : 0.7420   
## Neg Pred Value : 0.7896   
## Prevalence : 0.4145   
## Detection Rate : 0.2849   
## Detection Prevalence : 0.3839   
## Balanced Accuracy : 0.7590   
##   
## 'Positive' Class : 0   
##

# Plotting ROC Curve for Probit Model  
roc\_curve\_probit <- roc(testData$Depression, testData$predicted\_prob\_probit)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

plot(roc\_curve\_probit, main = "ROC Curve for Probit Model")



auc(roc\_curve\_probit)

## Area under the curve: 0.847

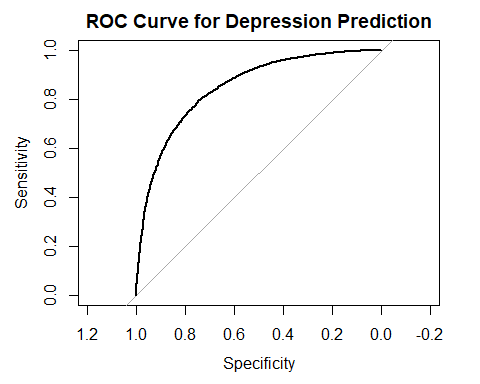
# Predict probabilities on the test set  
testData$predicted\_prob <- predict(model, newdata = testData, type = "response")  
  
# Thresholding to classify predictions  
testData$predicted\_class <- ifelse(testData$predicted\_prob > 0.5, 1, 0)  
  
# Confusion Matrix  
confusionMatrix(as.factor(testData$predicted\_class), testData$Depression)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 2377 818  
## 1 1092 4082  
##   
## Accuracy : 0.7718   
## 95% CI : (0.7626, 0.7807)  
## No Information Rate : 0.5855   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5243   
##   
## Mcnemar's Test P-Value : 4.194e-10   
##   
## Sensitivity : 0.6852   
## Specificity : 0.8331   
## Pos Pred Value : 0.7440   
## Neg Pred Value : 0.7889   
## Prevalence : 0.4145   
## Detection Rate : 0.2840   
## Detection Prevalence : 0.3818   
## Balanced Accuracy : 0.7591   
##   
## 'Positive' Class : 0   
##

# Plotting ROC Curve  
library(pROC)  
roc\_curve <- roc(testData$Depression, testData$predicted\_prob)

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

plot(roc\_curve, main = "ROC Curve for Depression Prediction")



auc(roc\_curve)

## Area under the curve: 0.847

# Calculate Pseudo R-squared  
pseudo\_r2 <- 1 - (model$deviance / null\_model$deviance)  
cat("\nPseudo R-squared:", pseudo\_r2, "\n")

##   
## Pseudo R-squared: 0.3063392

# Odds Probability Analysis  
# Display sample of predicted probabilities and odds from the logistic model  
testData$predicted\_odds <- testData$predicted\_prob / (1 - testData$predicted\_prob)  
print("Sample of Predicted Probabilities and Odds (Logistic Model):")

## [1] "Sample of Predicted Probabilities and Odds (Logistic Model):"

head(testData[, c("Depression", "predicted\_prob", "predicted\_odds")])

## Depression predicted\_prob predicted\_odds  
## 1 1 0.60510604 1.5323254  
## 2 0 0.20624668 0.2598372  
## 5 0 0.53273641 1.1401197  
## 6 0 0.11762833 0.1333093  
## 8 0 0.08262486 0.0900666  
## 9 1 0.64691334 1.8321659