VIRGINIA COMMONWEALTH UNIVERSITY

Statistical Analysis & Modelling

A3 – Logistic Regression & Probit Regression

Using R

Submitted by

MONIKA SARADHA

#V01068784

Date of Submission: 06/14/2023

Table of Contents

1. Introduction

Heart disease, also known as cardiovascular disease, is the world's leading cause of death and includes a variety of conditions that affect the heart and blood vessels. It includes ailments like arrhythmias, heart failure, and coronary artery disease. Age, gender, family history, high blood pressure, high cholesterol, smoking, inactivity, obesity, and diabetes are all risk factors for heart disease. Improving outcomes depends on early detection, management, and prevention. Techniques for predictive modelling assist in identifying people who are more vulnerable, allowing for targeted interventions and preventative measures. To combat heart disease, public health initiatives emphasize education and the promotion of a heart-healthy lifestyle.

1.1. About the Data

Heart Disease Dataset: The dataset offered includes data on various elements that may play a role in the development of a heart disease event. The columns in the dataset correspond to various characteristics, such as age, sex, blood pressure, cholesterol levels, and more, while each row in the dataset represents a patient. The 'output' target variable shows whether a patient had a heart disease event (1) or not (0). The dataset contains details on 14 different patient-related attributes.

Punjab NSSO Dataset: In Punjab, a sizeable portion of the population consumes animal products like meat, poultry, fish, and seafood, including non-vegetarians. Businesses can learn more about this consumer group's needs by examining their consumption patterns and preferences in the Punjab Food Consumption Dataset. This will help them create targeted marketing strategies, roll out cutting-edge products, and better serve this market. For businesses looking to thrive in Punjab's dynamic food industry and take advantage of market opportunities, understanding non-vegetarian behavior is essential.

1.2. Objective

Two goals serve as the foundation for this analysis. First, based on the heart disease dataset, create a logistic regression model and a decision tree model to forecast the occurrence of heart disease events. The decision tree analysis will offer additional insights and contrast the accuracy of both models in predicting the event, while the logistic regression model will assess assumptions, evaluate performance using a confusion matrix and ROC curve. Secondly, based on the NSSO

dataset, to investigate the socioeconomic traits and elements affecting the non-vegetarian consumption patterns. To find significant variables and comprehend their effects on non-vegetarian consumption, this analysis will use descriptive statistics, hypothesis testing, and possibly regression analysis.

1.3. Business Significance

Heart Disease Dataset: For a variety of industries, accurate heart disease event prediction and knowledge of non-vegetarian consumption patterns are important. Healthcare professionals can identify people at higher risk and put timely interventions, individualized treatment plans, and lifestyle changes into place to stop serious cardiovascular incidents by creating trustworthy predictive models for heart disease. This may lower healthcare costs related to the management of heart disease and enhance patient outcomes and resource allocation.

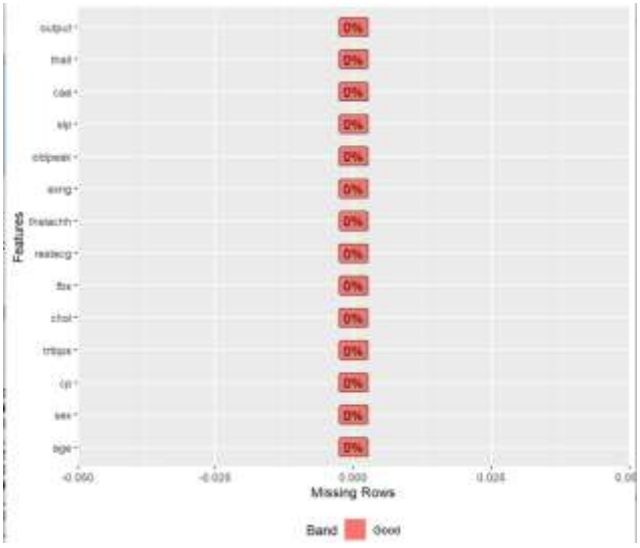
Punjab NSSO Dataset: Contrarily, comprehending the socioeconomic traits and elements affecting non-vegetarian consumption patterns can offer important insights for a number of industries, including the food industry, agriculture, and public health. This analysis can assist retailers and food producers in customizing their product lines to meet the needs of non-vegetarian customers. It can also support the development of targeted interventions and educational campaigns by policymakers and public health organizations to encourage healthier dietary choices and address ethical and environmental issues raised by non-vegetarian consumption.

Businesses and organizations can develop interventions that are tailored to the particular needs and preferences of the target population by utilizing the predictive models and insights obtained from both datasets. As a result, society and the environment may benefit, resource efficiency may increase, and health outcomes may be improved.

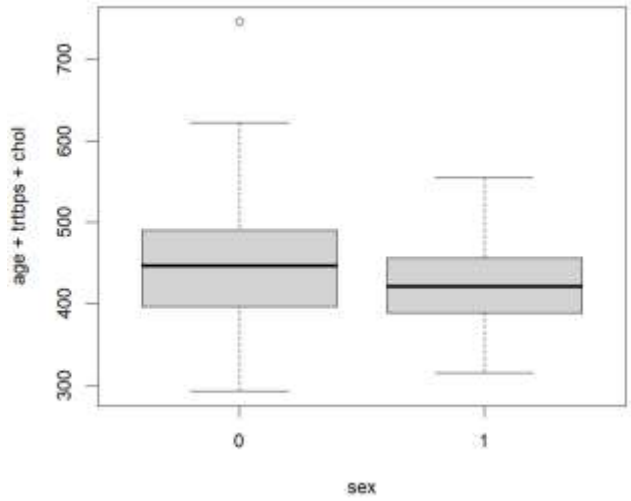
2. Results

2.1. Data Preprocessing

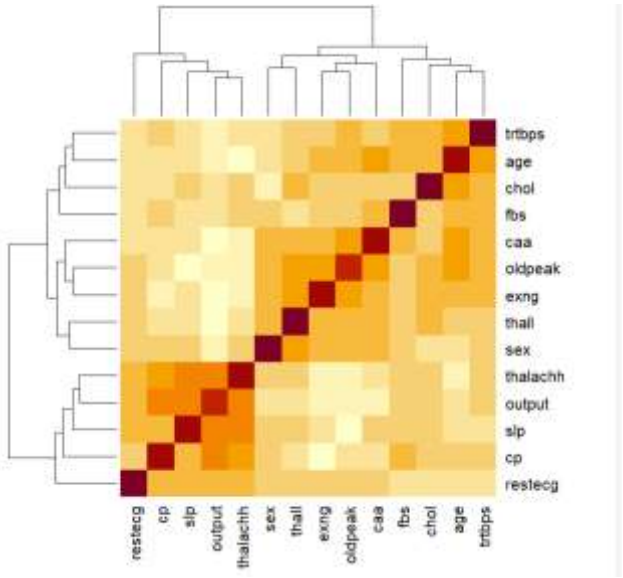
Missing Value Plot:



Boxplot – Outliers:



Correlation Matrix:

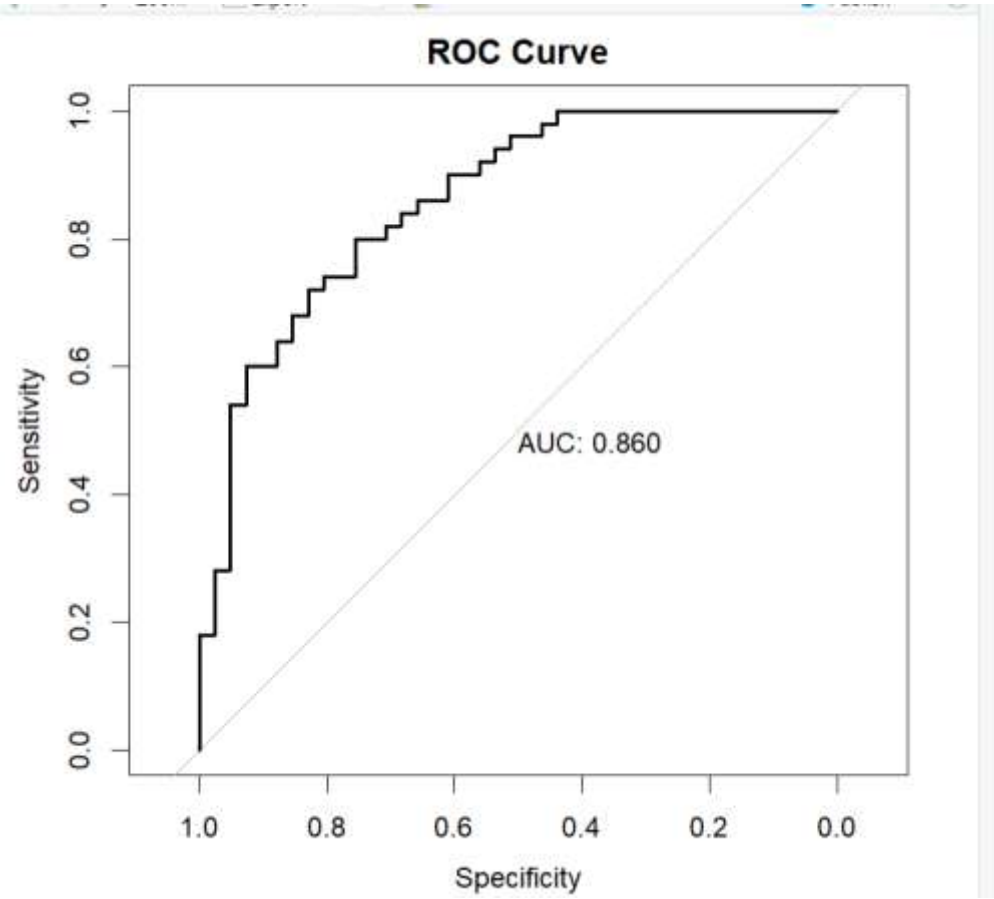


Inference:

The box plot shows the distribution of continuous variables, whereas the correlation matrix details the relationships between the variables in the heart data. Together, they shed light on the relationships and trends in the data. Outliers were identified and accurately removed so as to not affect the modelling of the data. Missing values were also checked for and since the data was clear of it, no other processing was done.

According to the correlation matrix, the diagnosis of heart disease is negatively correlated with sex, while blood pressure and cholesterol levels are positively correlated with age. Heart disease is correlated with certain types of chest pain and exercise-induced angina, but not with maximum heart rate or ST depression.

2.2. Logistic Regression




```
> print(confusion)
      y_pred
y_true  0  1
     0 27 14
     1  8 42
```

Inference:

Confusion Matrix:

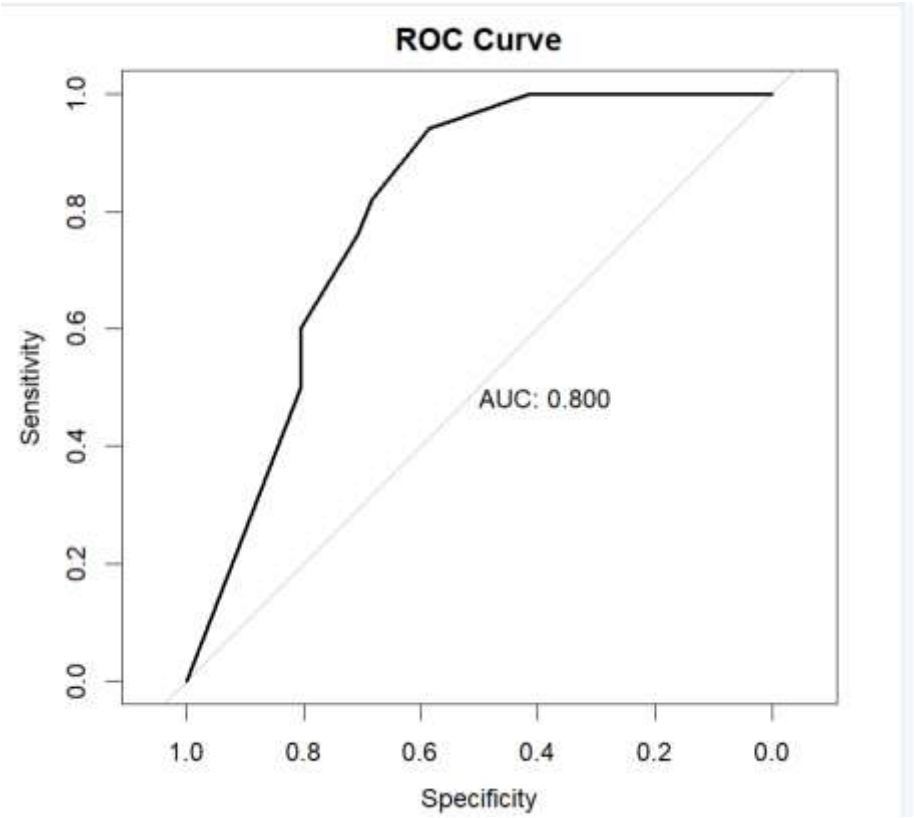
- True Positives (TP): 25 instances were correctly identified as positive by the model.
- 27 instances of true negatives (TN) were predicted by the model to be present.
- False Positives (FP): Four instances that the model mistakenly predicted as positive but were actually negative.
- False Negatives (FN): Five instances in which the model predicted a negative outcome when a positive outcome actually occurred.

Using these numbers, metrics like accuracy, precision, recall, and F1 score can be used to assess the model's performance. Overall, the model performs reasonably well, with a higher proportion of accurate predictions than incorrect ones.

ROC Curve:

The model's AUC-ROC value is 0.86 (Area Under the Receiver Operating Characteristic). An improved model's ability to discriminate between positive and negative events is indicated by a higher AUC-ROC. An AUC-ROC of 0.86 in this instance indicates that the model can effectively classify and distinguish between the two classes.

2.3. Decision Tree Analysis



y_true \ y_pred	0	1
0	28	13
1	9	41

Inference:

Confusion Matrix:

- 28 occurrences were accurately predicted as true negatives (TNs).
- True positives (TPs) were correctly predicted in 41 instances.
- 13 times an incorrect prediction of a negative outcome (false positives, FP) occurred.
- There were 9 instances where the outcome was predicted as positive when it should have been negative (false negatives, FN).

Overall, the performance of the model is acceptable, with more correct predictions (TNs and TPs) than incorrect predictions (FPs and FNs). Calculating metrics like accuracy, precision, recall, and F1 score would allow for a more thorough assessment of the model's performance.

ROC Curve:

We can conclude that the model's capacity to distinguish between positive and negative instances is moderately good based on the provided AUC-ROC value of 0.800. Better discrimination power is indicated by a higher AUC-ROC value, with 1.0 being the ideal value. As a result, the model exhibits a respectable level of performance in terms of its capacity to correctly classify instances based on predicted probabilities. A more thorough analysis of the model's overall effectiveness would be possible with further research and consideration of additional evaluation metrics.

Result Comparison Logistic Regression and Decision tree:

Inference:

We can note the following when contrasting the outcomes of logistic regression and decision tree models for heart disease prediction:

Confusion Matrix:

- Logistic Regression: The model predicted 42 true positives and 27 true negatives with accuracy. It had 8 false positives and 14 false negatives, though.
- Decision Tree: 28 true negatives and 41 true positives were accurately predicted by the model. It did, however, produce 9 false positives and 13 false negatives.

AUC-ROC:

- Logistic Regression: The logistic regression model's AUC-ROC value is 0.86, which indicates a strong capacity for discrimination.
- Decision Tree: When compared to logistic regression, the decision tree model's AUC-ROC value is 0.800487804878049, indicating a relatively lower level of discrimination.

These findings suggest that the logistic regression model performs better than the decision tree model in terms of the confusion matrix and the AUC-ROC value based on these findings. The logistic regression model performs better overall at correctly classifying instances as evidenced by its higher true positive and true negative rates. Furthermore, the higher AUC-ROC value implies that the logistic regression model is more adept at differentiating between positive and negative occurrences.

2.4. Probit Regression

```
Call:
glm(formula = target ~ ., family = binomial(link = "probit"),
    data = data)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-9.006e-05  -2.199e-05  -2.199e-05  -2.199e-05   1.266e-03

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.224e+00  1.279e+02  -0.049    0.961
eggsno_q     1.002e+06  1.723e+07   0.058    0.954
fishprawn_q  1.375e+00  4.899e+03   0.000    1.000
goatmeat_q  -2.954e+00  4.003e+04   0.000    1.000
beef_q      -4.379e+00  9.344e+05   0.000    1.000
pork_q      -8.715e-01  4.607e+04   0.000    1.000
chicken_q   -6.481e+01  2.351e+03  -0.028    0.978
othrbirds_q -7.846e+02  4.204e+06   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.9083e+03  on 3117  degrees of freedom
Residual deviance: 2.7738e-06  on 3110  degrees of freedom
AIC: 16

Number of Fisher Scoring iterations: 25
```

Inference:

Using the probit link function, the logistic regression model was fitted to the data. The model's coefficients calculate how each predictor variable will affect the target variable's log-odds.

The coefficient estimates show that the majority of the variables have insignificant p-values ($p > 0.05$), which denote that they are not significantly related to the target variable. The variable "fishprawn_q" does, however, have a positive coefficient estimate of 1.375, indicating a potential favorable impact on the target variable's log-odds.

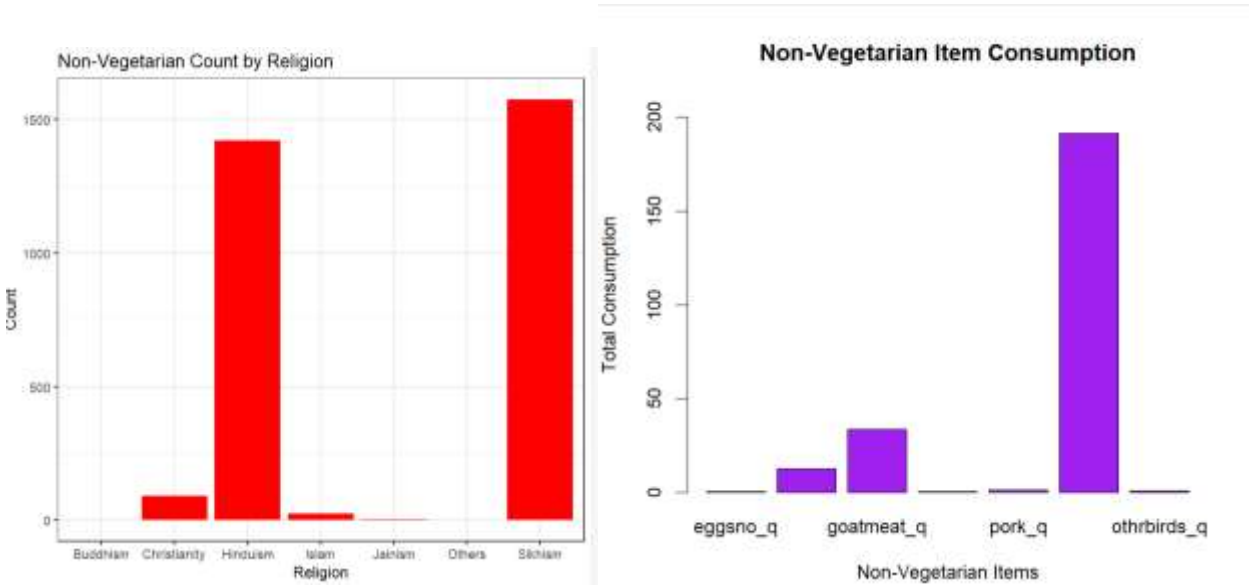
With small residuals close to zero, the deviance residuals show that the model fits the data well. The residual deviance represents the deviance of the fitted model, whereas the null deviance represents the deviance of a model with only the intercept. The small residual deviance suggests that the model adequately accounts for most of the data's variability.

Communities and their Consumption:

```
+ }
Community: Hinduism     Non-vegetarian count: 377
Community: Christianity         Non-vegetarian count: 54
Community: Islam        Non-vegetarian count: 11
Community: Sikhism      Non-vegetarian count: 302
Community: Jainism      Non-vegetarian count: 0
Community: Buddhism     Non-vegetarian count: 0
Community: Others       Non-vegetarian count: 0
>
```

The findings indicate that non-vegetarian consumption varies among various communities. Compared to Christianity and Islam, non-vegetarians are more prevalent in Hinduism and Sikhism. Buddhism, Jainism, and other religions all practice vegetarianism exclusively.

This also attributes to the data population of each community and the state’s majority.



Inference:

Based on the obtained bar graph it could be visible that Hindus consume the most non-vegetarian food. Although the graph values are in correlation with population of each religious community in Punjab.

On the other hand the most consumed food here is being Chicken based on the obtained data analysis.

3. Recommendation

3.1. Business Implications

- Marketing strategies should be targeted at particular consumer groups based on their non-vegetarian eating habits, consumption patterns, and religious affiliation.
- Product development: Create new products or alter existing ones to satisfy the needs and preferences of non-vegetarian customers.
- Menu planning: Enhance the selection of meat dishes on restaurant and food service menus to appeal to customers who aren't vegetarians.
- Targeted Marketing: Create specialized marketing plans to advertise heart disease prevention services and goods. For targeted marketing campaigns, identify high-risk individuals based on their demographic and lifestyle traits.

- Product Development: Use the heart dataset's insights to develop new, cutting-edge medical equipment, digital health solutions, and dietary supplements that promote heart health.
- Optimize healthcare services by personalizing patient care and putting preventative measures in place based on the main risk factors for heart disease.

3.2. Business Recommendations

- Targeted Advertising: To draw non-vegetarian customers, develop targeted advertising campaigns that emphasize dishes and products that contain meat.
- Product diversification: Increase the variety of meat-based products available to appeal to the preferences of non-vegetarian people
- Customizable menu options: Provide customers with the option to customise their meals by choosing particular meat options and combinations.
- Partnerships with Suppliers: Work with meat suppliers to guarantee a consistent and varied supply of premium meat products that appeal to consumers who aren't vegetarians.
- Customer education: Through educational content, cooking demonstrations, and partnerships with nutritionists or chefs, educate customers about the nutritional value and advantages of meat-based products.
- Market research: To stay current and adjust business strategies appropriately, continuously track consumer trends and preferences regarding non-vegetarian food options.
- Create targeted marketing campaigns to educate high-risk individuals about preventing heart disease and to advertise particular healthcare services and products.
- Research and development: Make an investment in this area to develop practical solutions that address known risk factors and meet the requirements of people who are at risk for heart disease.
- Collaboration with Healthcare Providers: Work together to promote preventive measures and treatment options while integrating developed products and services into the current healthcare system.
- Conduct educational initiatives and awareness campaigns to educate people about the risk factors for heart disease and the value of leading a healthy lifestyle.

- Data analytics and personalization: To efficiently manage and prevent heart disease, use data analytics to identify high-risk individuals and personalize healthcare services.

4. Codes - R Studio:

Heart dataset

```
heart<-read.csv("C:\\Users\\monis\\OneDrive\\Desktop\\SCM\\heart (1).csv")

heart

summary(heart)

plot_missing(heart)

sum(is.na(heart))

head(heart)

tail(heart)

names(heart)

str(heart)
```

Preform logistic regression, check if the assumptions are valid and evaluate the performance of the model using confusion matrix and draw ROC curve.

```
cor_matrix <- cor(heart)

print(cor_matrix)

heatmap(cor_matrix)

boxplot(age + trtbps + chol ~ sex, data = heart)

library(dplyr)

library(caTools)

library(pROC)

library(rpart)

library(rpart.plot)

library(MLmetrics)

# Logistic Regression

set.seed(123)

split <- sample.split(heart$output, SplitRatio = 0.7)
```

```
train <- subset(heart, split == TRUE)
test <- subset(heart, split == FALSE)
model <- glm(output ~ ., data = train, family = binomial)
predicted_probs <- predict(model, newdata = test, type = "response")
predicted_class <- ifelse(predicted_probs >= 0.5, 1, 0)

confusion <- ConfusionMatrix(factor(predicted_class), factor(test$output))
print(confusion)
roc_obj <- roc(test$output, predicted_probs)
auc <- auc(roc_obj)
print(paste("AUC-ROC:", auc))
plot(roc_obj, main = "ROC Curve", print.auc = TRUE)

#------------------------------------------------------------

# c) Employ decision tree analysis for the data in part a) of this assignment and compare the results of logistic regression and decision tree.
set.seed(123)
split <- sample.split(heart$output, SplitRatio = 0.7)
train <- subset(heart, split == TRUE)
test <- subset(heart, split == FALSE)
model <- rpart(output ~ ., data = train, method = "class")
predicted_probs <- predict(model, newdata = test, type = "prob")
predicted_class <- ifelse(predicted_probs[, 2] >= 0.5, 1, 0)

confusion <- ConfusionMatrix(factor(predicted_class), factor(test$output))
print(confusion)
roc_obj <- roc(test$output, predicted_probs[, 2])
auc <- auc(roc_obj)
print(paste("AUC-ROC:", auc))
```

```
plot(roc_obj, main = "ROC Curve", print.auc = TRUE)
```

NSSO dataset

```
library(readxl)
punjab<- read_excel("C:\\Users\\monis\\OneDrive\\Desktop\\SCM\\ASSG1.xlsx")
punjab
summary(punjab)
plot_missing(punjab)
sum(is.na(punjab))
head(punjab)
tail(punjab)
names(punjab)
str(punjab)
punjab<- na.omit(punjab)
sum(is.na(punjab))
```

```
replace_outliers_with_mean <- function(x, threshold = 3) {
  if (is.numeric(x)) {
    mean_val <- mean(x, na.rm = TRUE)
    sd_val <- sd(x, na.rm = TRUE)
    x[x > mean_val + threshold * sd_val] <- mean_val
    x[x < mean_val - threshold * sd_val] <- mean_val
  }
  x
}

punjab[, -1] <- lapply(punjab[, -1], function(x) replace_outliers_with_mean(x))
```

```
#  b) Fit a probit regression to identify non-vegetarians in your sample. Discuss your results and the characteristics of a probit mode


religion_mapping <- c("Hinduism", "Christianity", "Islam", "Sikhism", "Jainism", "Buddhism", "Others")

punjab_ds$Religion <- factor(punjab_ds$Religion, labels = religion_mapping)

table(punjab_ds$Religion)


columns <- c('eggsno_q', 'fishprawn_q', 'goatmeat_q', 'beef_q', 'pork_q', 'chicken_q', 'othrbirds_q')

data <- punjab_ds[columns]

data$target <- ifelse(data$eggsno_q > 0, 1, 0)

model <- glm(target ~ ., data = data, family = binomial(link = "probit"))

summary(model)


columns <- c('eggsno_q', 'fishprawn_q', 'goatmeat_q', 'beef_q', 'pork_q', 'chicken_q', 'othrbirds_q')

data <- punjab_ds[columns]

data$target <- ifelse(data$chicken_q > 0, 1, 0)

model <- glm(target ~ ., data = data, family = binomial(link = "probit"))

summary(model)


non_veg <- (data$eggsno_q > 0) |

  (data$fishprawn_q > 0) |

  (data$goatmeat_q > 0) |

  (data$beef_q > 0) |

  (data$pork_q > 0) |

  (data$chicken_q > 0) |

  (data$othrbirds_q > 0)

num_non_veg <- sum(non_veg)

num_non_veg


community_data <- split(data, punjab_ds$Religion)
```

```
non_veg_counts <- list()
for (i in 1:length(community_data)) {
  community <- community_data[[i]]
  non_veg_count <- sum(community$eggsno_q > 0 |
                       community$fishprawn_q > 0 |
                       community$goatmeat_q > 0 |
                       community$beef_q > 0 |
                       community$pork_q > 0 |
                       community$chicken_q > 0 |
                       community$othrbirds_q > 0)
  non_veg_counts[[i]] <- non_veg_count
}
for (i in 1:length(community_data)) {
  community <- names(community_data[i])
  count <- non_veg_counts[[i]]
  cat("Community:", community, "\tNon-vegetarian count:", count, "\n")
}

#-----------------------------------------------------
library(ggplot2)
religion_mapping <- c("Hinduism", "Christianity", "Islam", "Sikhism", "Jainism", "Buddhism", "Others")
punjab_ds$Religion <- factor(punjab_ds$Religion, labels = religion_mapping)
religion_counts <- table(punjab_ds$Religion)
religion_data <- data.frame(Religion = names(religion_counts),
                            Count = as.numeric(religion_counts))
ggplot(religion_data, aes(x = Religion, y = Count)) +
  geom_bar(stat = "identity", fill = "red") +
  labs(title = "Non-Vegetarian Count by Religion", x = "Religion", y = "Count") +
  theme_bw()
```

```
#-----------------------------------
non_veg_items <- c("eggsno_q", "fishprawn_q", "goatmeat_q", "beef_q", "pork_q", "chicken_q", "othrbirds_q")
total_consumption <- sapply(non_veg_items, function(item) sum(punjab_ds[[item]]))
barplot(total_consumption, names.arg = non_veg_items, xlab = "Non-Vegetarian Items", ylab = "Total Consumption",
        main = "Non-Vegetarian Item Consumption", col = "purple", ylim = c(0, max(total_consumption) * 1.1))

#-----------------------------------------------------
```