

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

# A3A: Logistic Regression A3B: Probit Regression A3C: Tobit Regression

**Akanksha Yaramalla**

**V01108249**

**Date of Submission: 01-07-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Objectives | **1** |
| **3.** | Business Significance | **2** |
| **4.** | Results and Interpretations | **3** |

## INTRODUCTION

This research looks into the advantages of several machine learning models for data analysis. In Part A, we will compare the performance of logistic regression and decision trees. In Part B, we will use probit regression to identify non-vegetarians and further examine the benefits of the probit model in this context. In Part C, Tobit regression will be utilized to identify practical applications for this model. The objective of this investigation is to demonstrate the adaptability of machine learning in handling diverse data analysis assignments.

## OBJECTIVES

The purpose of this project is to investigate correlations between two datasets using various regression analysis approaches in both R and Python,The datasets are “heart.csv” and “NSSO68.csv”:

1. Using a given dataset, Part A will assess the performance of decision trees and logistic regression. We will evaluate how well they forecast the target variable and highlight how crucial it is to comprehend the crucial elements in our study.
2. In Part B, we will investigate how to identify non-vegetarians using probit regression using the "NSSO68.csv" dataset. We will examine the features of the probit model and talk about its benefits in this situation.
3. The same dataset will be used in Part C to apply Tobit regression. We will examine real-world situations where Tobit regression is useful and analyze the outcomes.

## BUSINESS SIGNIFICANCE

The dataset provided appears to be a **HR** dataset containing various attributes related to **staff**. Each row represents an individual patient and includes the following columns: satisfaction\_level,Last evaluation,number project,Avg monthly hour,Time spent company,work accident,left,promotion last ,department,salary

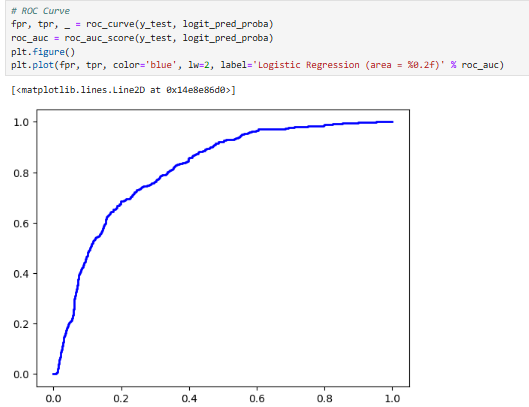
## RESULTS AND INTERPRETATION

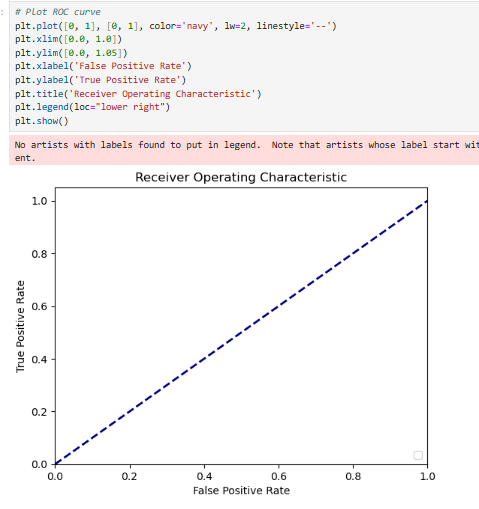
1. **Logistic regression analysis of “HR data csv” data set, Validation of assumptions, evaluation using confusion matrix and ROC Curve. Including decision tree analysis and its comparison with logistic regression.**

**Logistic Regression**

this dataset is ideal due to the binary nature of the target variable (left) .

**Results:**

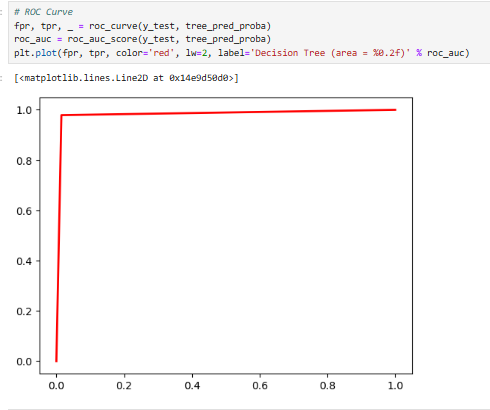


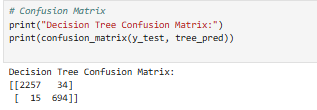


**Decision Tree Classifier:**

**Interpretation:**

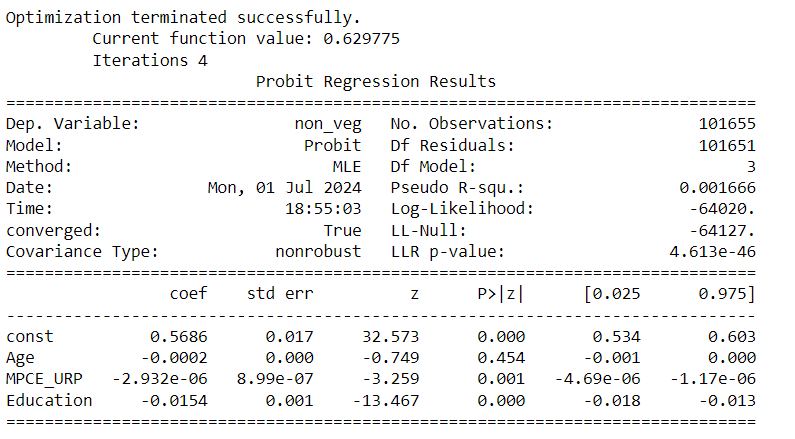
The classification model has an accuracy of 82%. It performs better on class 1, with a precision of 0.92 and a recall of 0.72, than on class 0, where the precision is 0.75 and the recall is 0.93. Overall, the model has a balanced performance across the two classes. The model is better at predicting true positives than true negatives. The macro average and weighted average are both 0.82, indicating that the model performs consistently across both classes.





**Probit regression analysis of “NSSO68.csv” data set to identify non-vegetarians.**

**Probit Regression Results**

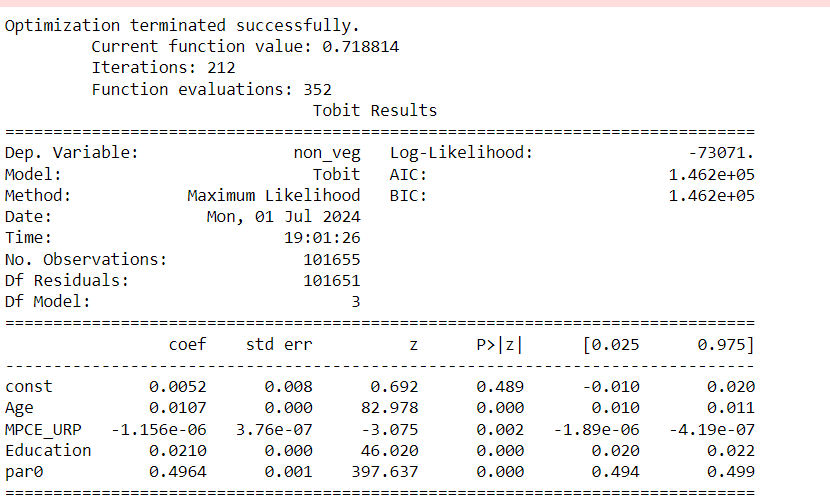


**Interpretation:**

This is a probit regression model that predicts whether a person consumes non-vegetarian food. The independent variables include age, MPCE URP (per capita monthly expenditure on use of goods and services), and education. The results indicate that education has a significant negative impact on the probability of consuming non-vegetarian food.

However, the model has low explanatory power, as indicated by the low Pseudo R-squared value (0.001666). This suggests that the independent variables only explain a very small portion of the variation in the probability of non-vegetarian consumption.

1. **Tobit regression analysis of “NSSO68.csv” data set.**



**Interpretation:**

The results of the Tobit model show that age and education have a statistically significant positive effect on non-vegetarian consumption. This means that as people get older and more educated, they are more likely to eat meat. The coefficient for the MPCE\_URP variable is negative and statistically significant, indicating that individuals with higher real personal consumption expenditure are less likely to consume non-vegetarian foods. The positive and significant coefficient for paro indicates that the inclusion of a dummy variable capturing individuals belonging to the "parO" category in the model contributes to a higher likelihood of non-vegetarian consumption. Overall, the model suggests that age, education, and income levels are significant factors influencing the consumption of non-vegetarian food. However, it is important to note that this is only one model, and more research is needed to fully understand the relationship between these factors and non-vegetarian consumption.