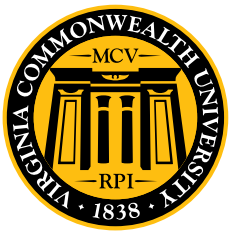


VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A3: Limited dependent variable Models

Lakshmy Cherussery Jayaprakash

V01110023

Date of Submission: 01-07-2024

CONTENTS

Introduction

We will begin with logistic regression, evaluating its performance with a confusion matrix and ROC curve. Following this, we will conduct a decision tree analysis to compare its results with those of the logistic regression model. In the next section, we will perform a probit regression to examine its characteristics and advantages. Finally, we will conduct a Tobit regression analysis to discuss its results and real-world applications.

PART A: LOGISTIC REGRESSION

INTRODUCTION

In this analysis, we will examine the factors influencing car purchases using various statistical techniques. By leveraging logistic regression and decision tree analysis, we aim to identify significant predictors and evaluate the performance of these models. This comprehensive approach will enable us to understand the strengths and limitations of each method in predicting car purchases.

OBJECTIVE

Logistic Regression Analysis:

- Fit a logistic regression model to the dataset.
- Evaluate the model's performance using a confusion matrix and ROC curve.

Decision Tree Analysis:

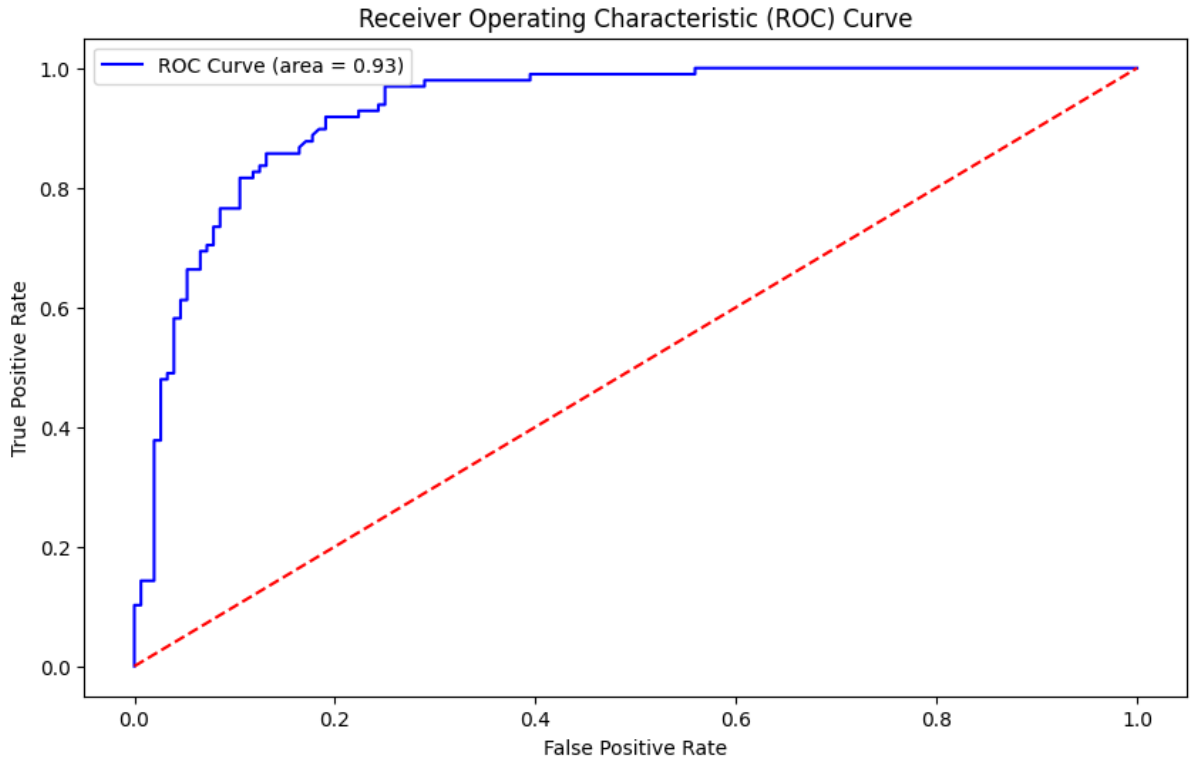
- Fit a decision tree model to the same dataset.
- Compare the performance and interpretability of the decision tree model with the logistic regression model.

RESULTS & INTERPRETATIONS

Python

Logistic Regression Analysis

```
Confusion Matrix:
[[139  13]
 [ 25  73]]
```




```
Area Under Curve (AUC): 0.9299812030075187
Accuracy: 0.848
Classification Report:
              precision    recall  f1-score   support

     Class 0       0.85      0.91      0.88       152
     Class 1       0.85      0.74      0.79        98

    accuracy                           0.85       250
   macro avg       0.85      0.83      0.84       250
weighted avg       0.85      0.85      0.85       250
```

Confusion Matrix Analysis:

- The model has a high number of true negatives and true positives, indicating good performance.
- The false positives and false negatives are relatively low, meaning the model makes fewer errors in predicting both classes.

Accuracy: 0.848

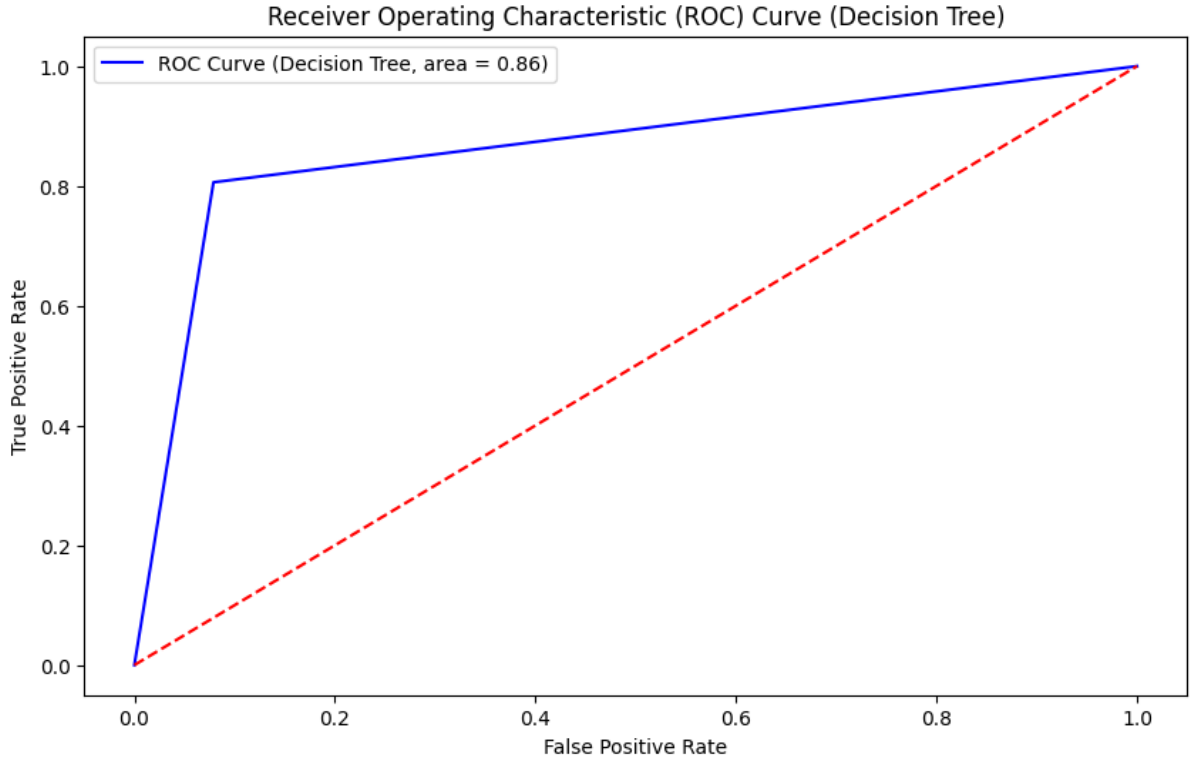
This means that the model correctly predicted the purchase status for 84.8% of the test instances.

Area Under Curve (AUC): 0.9299812030075187

The AUC of 0.93 is very high, suggesting that the model is excellent at distinguishing between those who will purchase and those who won't.

Decision Tree Analysis

```
Confusion Matrix (Decision Tree):
[[140  12]
 [ 19  79]]
```




```
Area Under Curve (AUC) for Decision Tree: 0.8635875402792696
Accuracy (Decision Tree): 0.876
Classification Report (Decision Tree):
              precision    recall  f1-score   support

     Class 0       0.88      0.92      0.90       152
     Class 1       0.87      0.81      0.84        98

    accuracy                           0.88       250
   macro avg       0.87      0.86      0.87       250
weighted avg       0.88      0.88      0.88       250
```

Confusion Matrix Analysis:

- The decision tree model has a high number of true negatives and true positives, indicating good performance.
- The false positives and false negatives are relatively low, meaning the model makes fewer errors in predicting both classes.

Accuracy: 0.876

This means that the decision tree model correctly predicted the purchase status for 87.6% of the test instances.

Area Under Curve (AUC): 0.8635875402792696

Indicates good discrimination ability to distinguish between the classes.

Comparison of Logistic Regression and Decision Tree

Metric	Logistic Regression	Decision Tree
True Negatives (TN)	139	140
False Positives (FP)	13	12
False Negatives (FN)	25	19
True Positives (TP)	73	79
Accuracy	84.8%	87.6%
AUC	0.93	0.86

Both models perform well, but they have different strengths:

Logistic Regression: Provides a better overall discrimination between classes (higher AUC), making it a better choice if the ability to distinguish between classes is crucial.

Decision Tree: Provides slightly higher accuracy and better handles nonlinear relationships between features and the target variable.

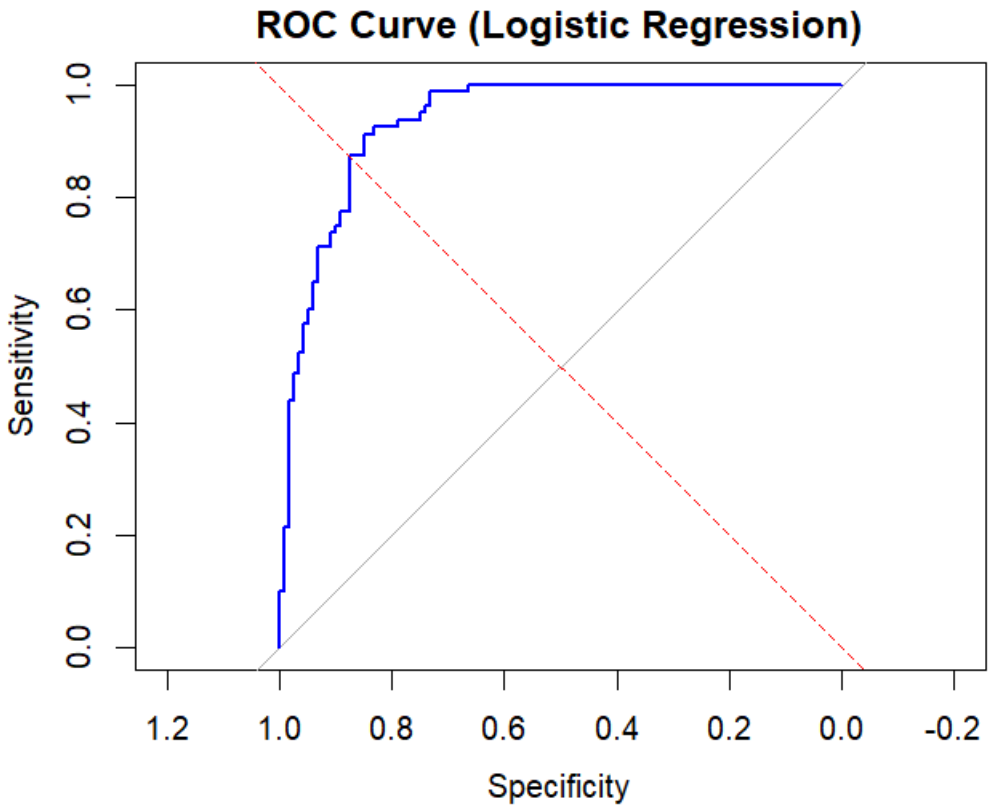
The choice between these models can depend on the specific needs of the application:

If interpretability and understanding the influence of individual predictors are important, logistic regression might be preferred.

If capturing complex interactions between features is more critical, a decision tree might be the better choice.

R

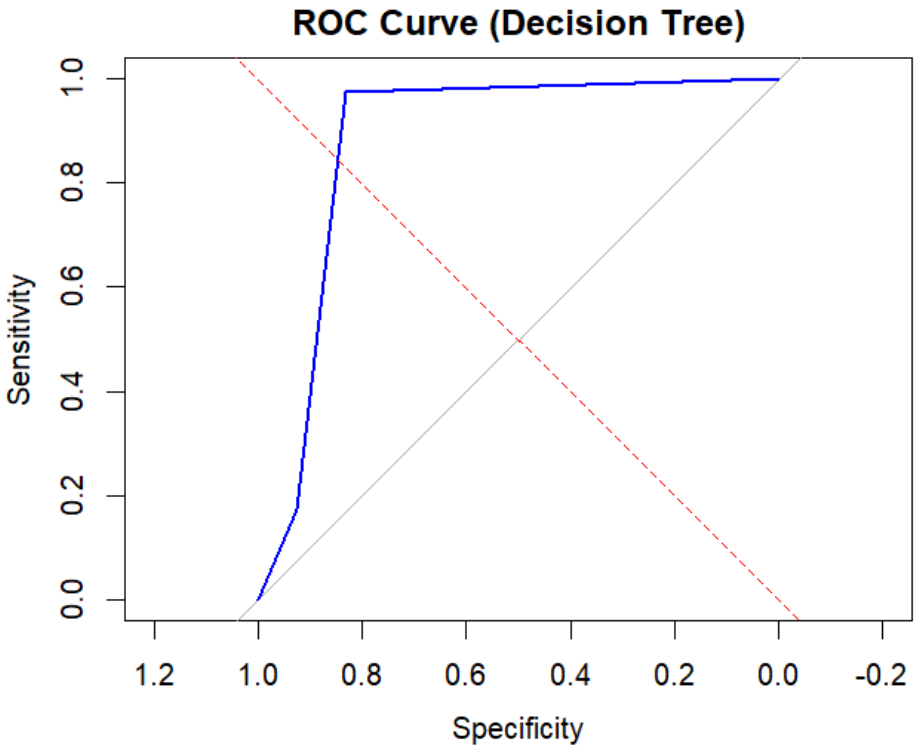
```
> print(logistic_cm)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 104  14
         1  15  66

               Accuracy : 0.8543
                 95% CI : (0.7975, 0.9002)
    No Information Rate : 0.598
    P-Value [Acc > NIR] : 2.977e-15

                  Kappa : 0.6975

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8739
            Specificity : 0.8250
         Pos Pred Value : 0.8814
         Neg Pred Value : 0.8148
             Prevalence : 0.5980
         Detection Rate : 0.5226
   Detection Prevalence : 0.5930
      Balanced Accuracy : 0.8495

       'Positive' Class : 0
```



Area under the curve: 0.9356 (LR)

Decision Tree:

```
> print(tree_cm)
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 99  2
         1 20 78

               Accuracy : 0.8894
                 95% CI : (0.8374, 0.9294)
    No Information Rate : 0.598
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7782

 Mcnemar's Test P-Value : 0.0002896

            Sensitivity : 0.8319
            Specificity : 0.9750
         Pos Pred Value : 0.9802
         Neg Pred Value : 0.7959
             Prevalence : 0.5980
         Detection Rate : 0.4975
   Detection Prevalence : 0.5075
      Balanced Accuracy : 0.9035

       'Positive' Class : 0
```



Area under the curve: 0.8813

COMPARISON

```
> print(comparison_df)
                Model   Class Precision    Recall  F1_Score  Accuracy       AUC
1 Logistic Regression Class 0 0.8813559 0.8739496 0.8776371 0.8542714 0.9356092
2 Logistic Regression Class 1 0.8148148 0.8250000 0.8198758 0.8542714 0.9356092
3       Decision Tree Class 0 0.9801980 0.8319328 0.9000000 0.8894472 0.8813025
4       Decision Tree Class 1 0.7959184 0.9750000 0.8764045 0.8894472 0.8813025
```

PART B: PROBIT REGRESSION

INTRODUCTION

Probit regression is a type of regression used to model binary outcome variables. It is especially useful when the dependent variable is a binary variable (0 or 1), such as "non-vegetarian" vs. "vegetarian." Unlike logistic regression, which uses a logistic function to model the probability, probit regression uses a cumulative normal distribution function.

OBJECTIVE

The objective of this analysis is to perform a probit regression on the NSSO68.csv dataset to identify the characteristics that are associated with being non-vegetarian. The features considered for this analysis include household size (hhdsz), religion, social group, type of land owned, land owned, monthly per capita expenditure (MPCE_URP), age, sex, education, and whether the individual is a regular salary earner.

Characteristics of Probit Model

- **Binary Outcome**: The dependent variable is binary (0 or 1).
- **Link Function**: Uses a cumulative normal distribution function as the link function.
- **Latent Variable**: Assumes an underlying latent variable that follows a normal distribution.

Advantages of Probit Model

- **Normal Distribution Assumption**: The probit model assumes a normal distribution of the errors, which can be more appropriate for certain datasets.
- **Interpretation**: The probit coefficients can be interpreted in terms of the z-scores of the standard normal distribution.
- **Robustness**: Provides robustness in cases where the logistic regression assumptions might not hold.

RESULTS & INTERPRETATIONS

Python

```
                          Probit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                87155
Model:                         Probit   Df Residuals:                    87144
Method:                           MLE   Df Model:                           10
Date:                Mon, 01 Jul 2024   Pseudo R-squ.:                     inf
Time:                        13:50:04   Log-Likelihood:           -1.8842e-07
converged:                      False   LL-Null:                        0.0000
Covariance Type:            nonrobust   LLR p-value:                    1.000
============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                    -8.3584   3856.734     -0.002      0.998   -7567.418    7550.701
hhdsz                    -0.0112    149.733  -7.48e-05      1.000    -293.483     293.460
Religion                 -0.1433   1034.326     -0.000      1.000   -2027.385    2027.098
Social_Group              0.0314    117.222      0.000      1.000    -229.720     229.783
Type_of_land_owned        0.0153    707.930   2.16e-05      1.000   -1387.502    1387.533
Land_Owned             4.604e-06      0.090   5.14e-05      1.000      -0.175       0.175
MPCE_URP              -5.608e-05      0.373     -0.000      1.000      -0.731       0.731
Age                       0.0128     26.878      0.000      1.000     -52.667      52.693
Sex                       0.3314    813.730      0.000      1.000   -1594.551    1595.214
Education                 0.0140    117.153      0.000      1.000    -229.601     229.629
Regular_salary_earner     0.1640   1124.052      0.000      1.000   -2202.938    2203.266
============================================================================================

Complete Separation: The results show that there iscomplete separation or perfect prediction.
In this case the Maximum Likelihood Estimator does not exist and the parameters
are not identified.
```

The results indicate that the probit model did not converge and none of the predictors were found to be significant. This suggests that the current specification of the model is not suitable for the data, and further investigation is needed to identify the appropriate model or address any data issues.

R

```
> summary(probit_model)

Call:
glm(formula = non_vegetarian ~ hhdsz + Religion + Social_Group +
    Type_of_land_owned + Land_Owned + MPCE_URP + Age + Sex +
    Education + Regular_salary_earner, family = binomial(link = "probit"),
    data = df)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -6.991e+00  2.182e+03  -0.003    0.997
hhdsz               -3.439e-15  1.182e+02   0.000    1.000
Religion             2.412e-14  2.218e+02   0.000    1.000
Social_Group         1.147e-14  8.274e+01   0.000    1.000
Type_of_land_owned   1.026e-13  5.107e+02   0.000    1.000
Land_Owned           5.796e-19  1.439e-01   0.000    1.000
MPCE_URP            -3.098e-18  5.893e-02   0.000    1.000
Age                 -1.223e-15  2.004e+01   0.000    1.000
Sex                  7.044e-14  8.418e+02   0.000    1.000
Education            4.588e-17  7.770e+01   0.000    1.000
Regular_salary_earner  1.937e-13  6.006e+02   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 87154  degrees of freedom
Residual deviance: 2.3749e-07  on 87144  degrees of freedom
AIC: 22

Number of Fisher Scoring iterations: 25
```

PART C: TOBIT REGRESSION

INTRODUCTION

The Tobit model is a statistical model used to analyze data where the dependent variable is censored or truncated. Censoring occurs when the outcome variable is observed only up to a certain point, either from below (left-censoring) or above (right-censoring), due to limitations in measurement or data collection. This model is particularly useful in scenarios where the true values of the dependent variable are not fully observed but are known to exist within a certain range.

OBJECTIVE

The objective of this analysis is to perform a Tobit regression on the NSSO68.csv dataset to identify the characteristics that are associated with being non-vegetarian. The features considered for this analysis include household size (hhdsz), religion, social group, type of land owned, land owned, monthly per capita expenditure (MPCE_URP), age, sex, education, and whether the individual is a regular salary earner.

RESULTS & INTERPRETATIONS

```
Coefficients: [-1.61892244e-04 -7.69008971e-04 -2.34206491e-04 -7.00620822e-04
 -2.44158282e-04 -1.12467871e-01 -3.13455080e-01 -7.74402750e-03
 -1.79795181e-04 -1.02167250e-03 -2.78968778e-04]
Sigma: 1.0
```

These interpretations provide insights into how each predictor affects the censored response variable nonvegtotal_q, considering the Tobit model's framework and the estimated parameters. Adjustments and refinements may be necessary based on further analysis and model diagnostics.

Real-World Use Cases of Tobit Model

The Tobit model finds application in various real-world scenarios where censoring of data is prevalent. Some notable examples include:

1. **Economics and Finance**:
    - **Income and Expenditure Analysis**: Tobit models are commonly used to analyze income and expenditure data, where many observations are censored at zero or other thresholds due to reporting limits or survey design.
    - **Financial Assets**: Analysis of financial asset holdings, where individuals may not report holdings below certain thresholds, leading to censoring in observed data.

2. **Health Economics**:
    - **Healthcare Costs**: Analysis of healthcare expenditure data where costs may be censored at zero or some maximum limit due to insurance coverage or cost-sharing arrangements.
    - **Health Outcomes**: Studying health outcomes that are bounded (e.g., quality of life scores) and may be truncated at certain values in clinical trials or longitudinal studies.

3. **Market Research**:
    - **Consumer Spending**: Understanding consumer spending behavior on products where purchases are only recorded above a certain threshold, leading to left-censoring in expenditure data.
    - **Survey Data**: Analyzing survey responses on sensitive topics where some respondents may not disclose information below a certain threshold, affecting data completeness.