VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A3: Limited dependent variable Models

MICAH ASHADEEP EMMANUEL

V01101166

Date of Submission: 01-07-2024

CONTENTS

Introduction

This analysis aims to comprehensively evaluate various statistical models to understand their applications and significance in real-world data analysis. Specifically, we will:

Part A: Conduct logistic regression and decision tree analysis on a car evaluation dataset to compare their performance and interpret the results.

Part B: Perform a probit regression to identify non-vegetarians in the NSSO68 dataset, discussing the characteristics and advantages of the probit model.

Part C: Conduct a Tobit regression analysis on the NSSO68 dataset, interpreting the results and exploring the real-world use cases of the Tobit model.

Objectives

1. **Part A:**
    - Validate assumptions of logistic regression.
    - Evaluate the model with a confusion matrix and ROC curve.
    - Compare the performance of logistic regression with a decision tree model.
2. **Part B:**
    - Conduct a probit regression to identify non-vegetarians.
    - Discuss the characteristics and advantages of the probit model.
3. **Part C:**
    - Perform a Tobit regression analysis.
    - Discuss the results and real-world applications of the Tobit model.

Business Significance

Understanding these statistical models is crucial for making informed business decisions. Logistic and probit regressions are often used in binary classification problems, helping businesses segment customers, predict churn, or detect fraud. Decision trees provide an intuitive way to make complex decisions based on various criteria. Tobit regression is used when dealing with censored data, such as in cases where the variable of interest has a limited range. By mastering these models, businesses can better analyze data, derive insights, and make data-driven decisions.

Part A: Logistic Regression and Decision Tree Analysis

Introduction

In the realm of predictive modelling, classification tasks play a crucial role in various domains, from medical diagnosis to customer segmentation. Two widely used techniques for binary classification are Logistic Regression and Decision Tree analysis. This part of the study aims to leverage these techniques to analyze a dataset from the UCI Machine Learning Repository, specifically focusing on the "Car Evaluation" dataset. The primary goal is to predict whether a car evaluation is acceptable or unacceptable based on various features.

Objectives

1. **Conduct a Logistic Regression Analysis**:
   - Validate the assumptions of logistic regression.
   - Evaluate model performance using a confusion matrix and ROC curve.
   - Interpret the results in terms of model coefficients and predictive power.
2. **Perform a Decision Tree Analysis**:
   - Build a decision tree model to classify car evaluations.
   - Compare its performance with the logistic regression model using similar evaluation metrics.
3. **Comparative Analysis**:
   - Assess and compare the precision, recall, F1 score, accuracy, and other relevant metrics for both models.
   - Discuss the interpretability, strengths, and recommendations of each approach.

Business Significance

Understanding the factors that contribute to a car being evaluated as acceptable or unacceptable has significant implications for automobile manufacturers and dealers. By accurately predicting the acceptability of cars based on their features, businesses can:

Improve Product Design: Insights from the models can guide manufacturers in designing cars that meet consumer expectations and regulatory standards.

Enhance Marketing Strategies: By identifying key features that influence acceptability, marketers can tailor their campaigns to highlight these attributes.

Optimize Inventory Management: Dealers can better manage their inventory by understanding the likelihood of cars being evaluated as acceptable, thereby reducing unsold stock and increasing customer satisfaction.

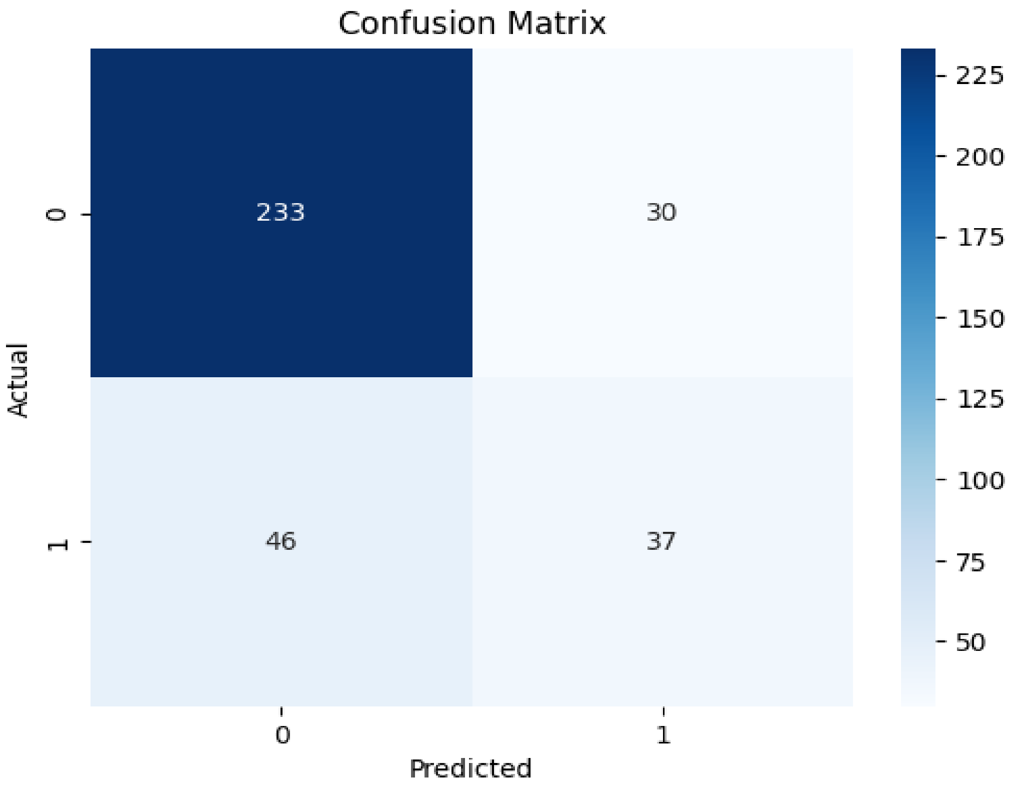
RESULTS

USING PYTHON:

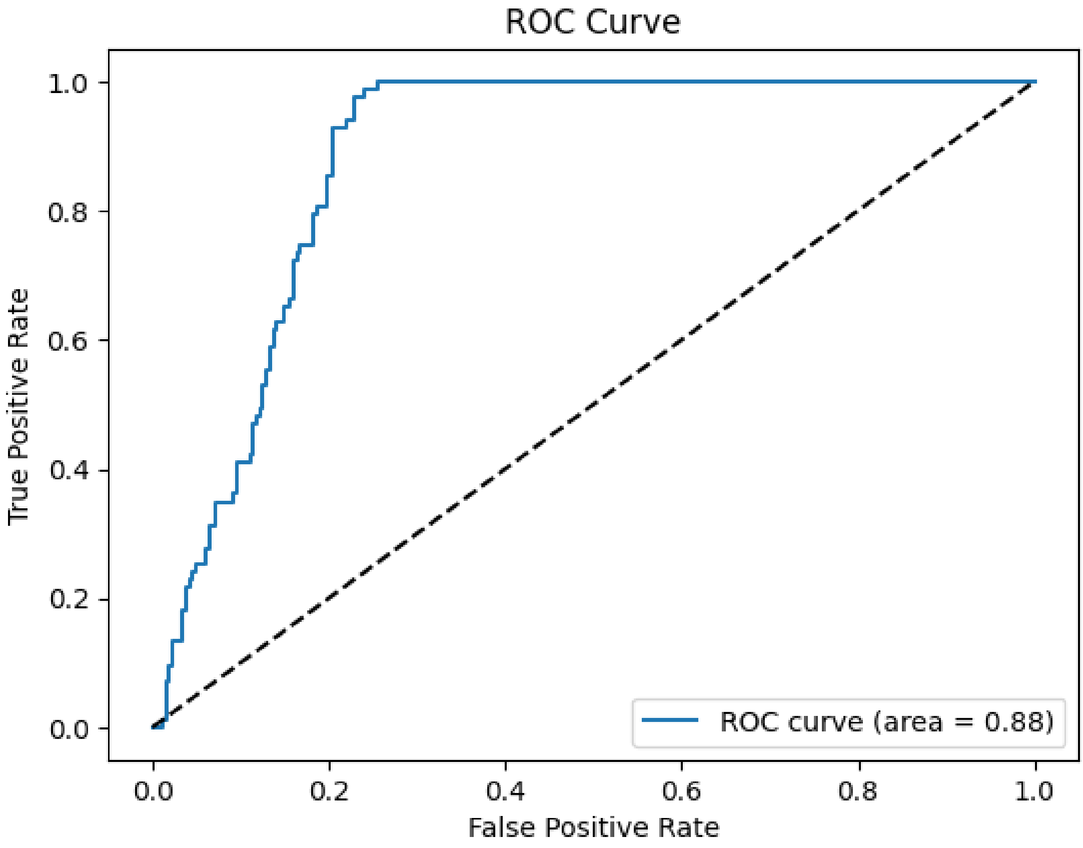
```
# Display the coefficients
buying price_low -0.437091
buying price_med 0.052327
buying price_v high -0.753111
maintenance cost_low -0.318367
maintenance cost_med  0.355681
maintenance cost_vhigh -0.631895
number of doors_3  0.482751
number of doors_4  0.613744
number of doors_5more  0.605615
number of persons_4  3.74772
number of persons_more 3.725572
 luggage boot_med_ -0.072464
 luggage boot_small  -0.620503
safety_low  -4.50456
safety_med -0.36074
```

```
# Confusion matrix
```

The picture below shows the Confusion Matrix.


Confusion Matrix
Actual
0
1
233
30
46
37
0
1
Predicted
225
200
175
150
125
100
75
50

#ROC Curve

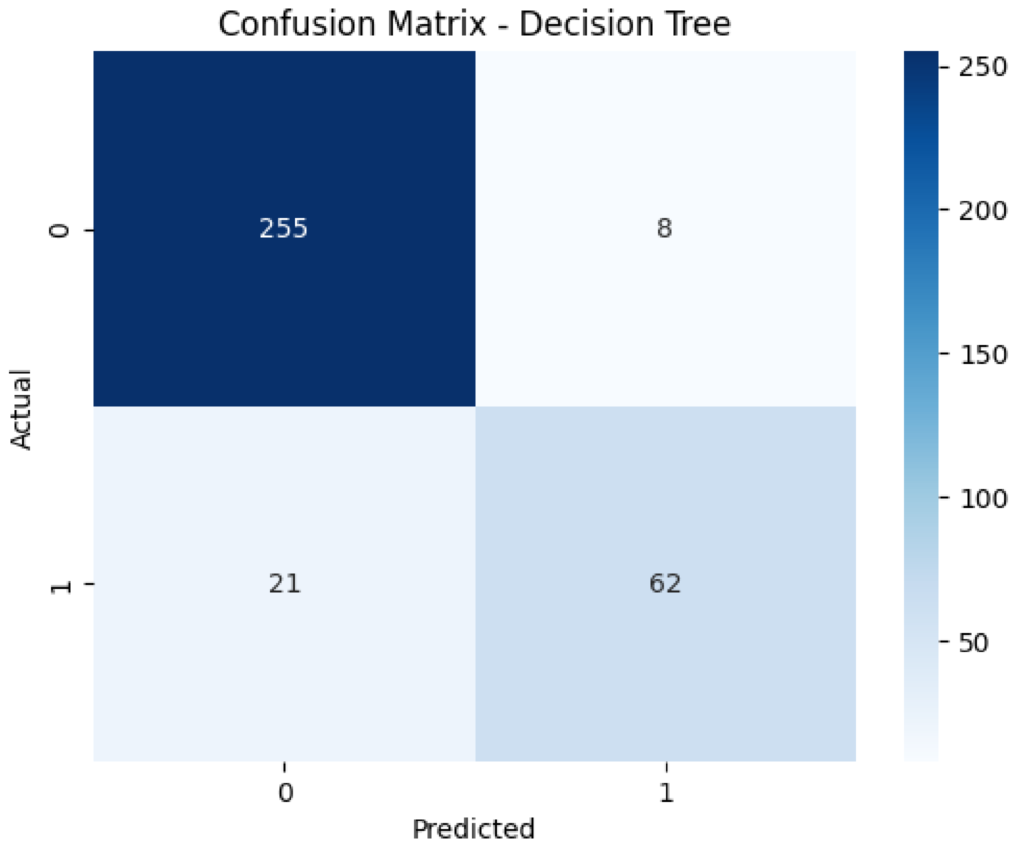

ROC Curve
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.88)

Accuracy(Logistic Regression)	0.78
Precision (Logistic Regression)	0.55
Recall (Logistic Regression)	0.45
F1 Score (Logistic Regression)	0.49

The above table shows Accuracy, Precision, Recall and F1 score

# Confusion matrix for decision tree

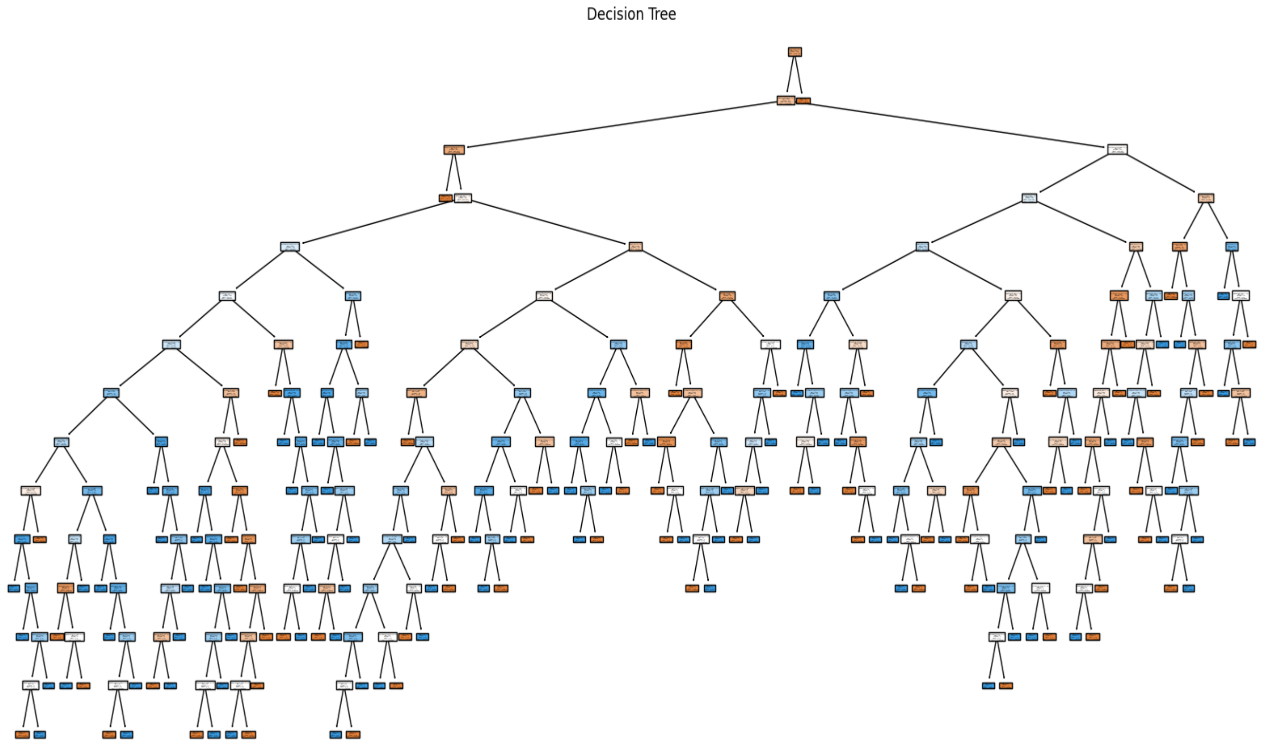
The below diagram shows the confusion matric for decision tree



# Accuracy for decision tree

Accuracy (Decision Tree): 0.92

# Plot the decision tree



COMPARISON OF ACCURACY BETWEEN LOGISTIC REGRESSION AND DECISION TREE

Logistic Regression	0.78
Decision Tree	0.92

The decision tree model has a significantly higher accuracy (92%) compared to the logistic regression model (78%). This suggests that the decision tree is more effective at correctly classifying car evaluations in this dataset.

USING R:

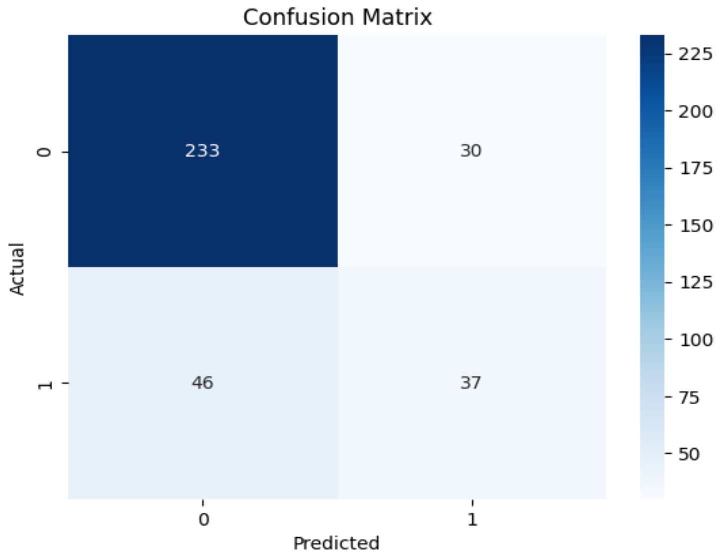
RESULTS

```
Coefficients:
                         Estimate   Std.Error  zvalue    Pr(>|z|)
(Intercept)             -19.52918  711.73869  -0.027    0.978110
buying.pricelow          -0.63058    0.23940  -2.634    0.008440
buying.pricemed           0.09850    0.24095   0.409    0.682692
buying.pricevhigh        -0.90627    0.24318  -3.727    0.000194
maintenance.costlow      -0.37164    0.24022  -1.547    0.121845
maintenance.costmed       0.17690    0.24064   0.735    0.462267
maintenance.costvhigh    -1.00190    0.24526  -4.085    4.41e-05
number.of.doors3          0.42001    0.24076   1.745    0.081068
number.of.doors4          0.62418    0.24536   2.544    0.010963
number.of.doors5more      0.45858    0.24080   1.904    0.056854
number.of.persons4       20.19501  711.73865   0.028    0.977364
number.of.personsmore    20.13746  711.73864   0.028    0.977428
luggage.bootmed          -0.08757    0.20883  -0.419    0.674968
luggage.bootsmall        -0.53830    0.21274  -2.530    0.011397
safetylow               -20.26051  715.92023  -0.028    0.977423
safetymed                -0.24805    0.17145  -1.447    0.147965
```

#Confusion matrix

```
> print(conf_matrix)

         Actual
Predicted   0   1
        0 233  35
        1  40  38
```

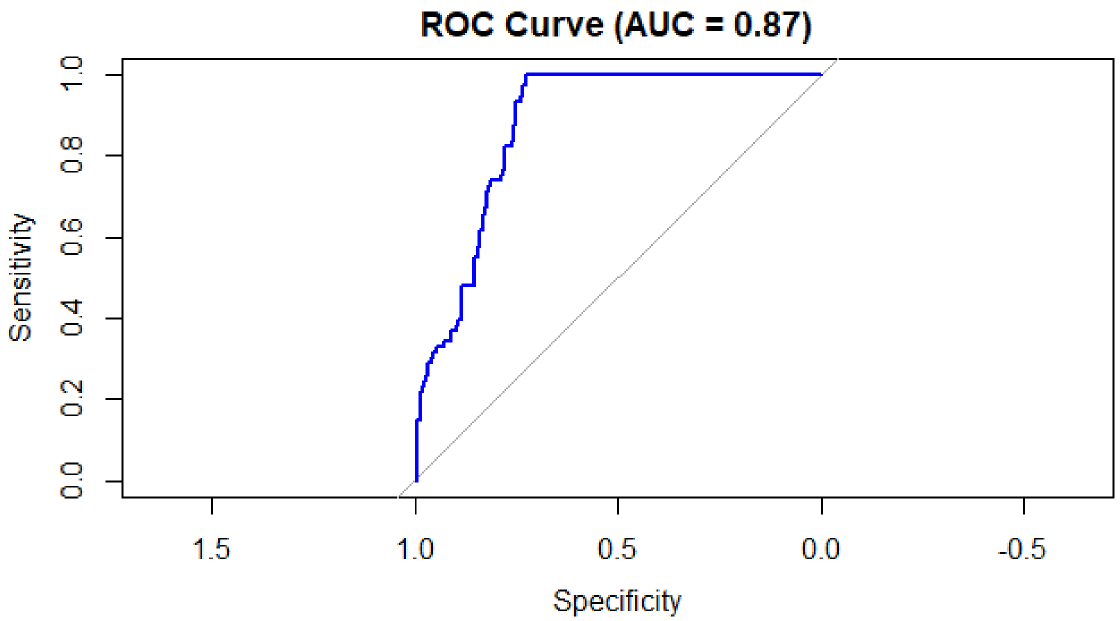


The picture is to show the similarity of confusion matrix derived from python and r

#ROC Curve

AUC=0.87

The picture below shows the ROC Curve.



Accuracy(Logistic Regression)	0.78
Precision (Logistic Regression)	0.55
Recall (Logistic Regression)	0.45
F1 Score (Logistic Regression)	0.49

Confusion Matrix of decision tree

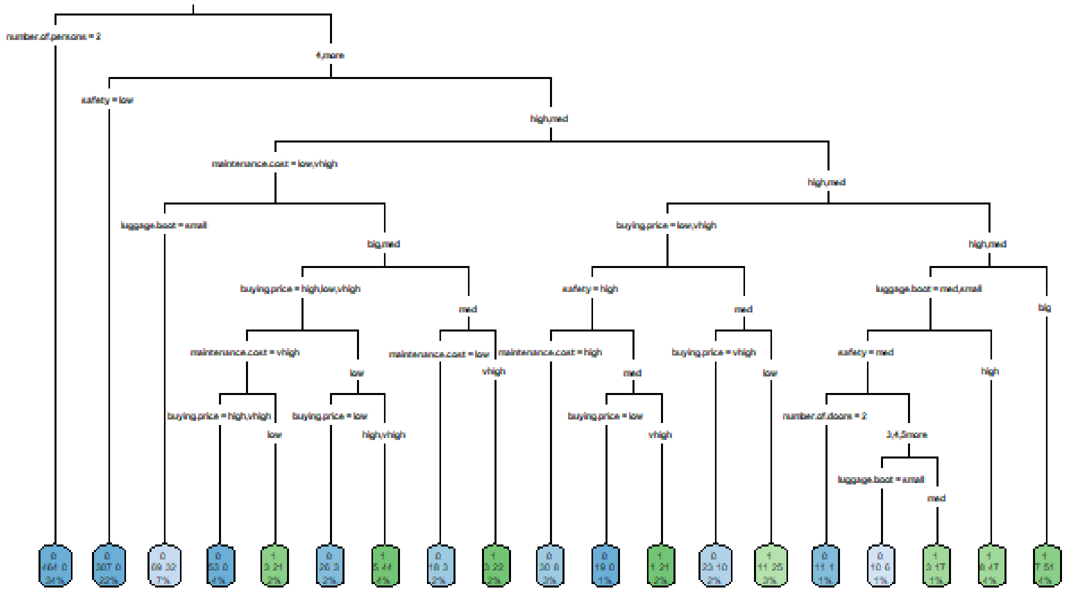
```
> print(tree_conf_matrix)
```

```
          Actual
Predicted   0   1
        0 266  17
        1   7  56
```

Decision Tree

#plot the decision tree

The picture below shows the decision tree.



Decision Tree Accuracy: 0.9306358
Decision Tree Precision: 0.8888889
Decision Tree Recall: 0.7671233
Decision Tree F1 Score: 0.8235294

COMPARISON BETWEEN LOGISTIC REGRESSION AND DECISION TREE

Logistic Regression a) Accuracy b) Precision c) Recall d) F1 Score	 0.78 0.55 0.45 0.49
Decision Tree a) Accuracy b) Precision c) Recall d) F1 Score	 0.93 0.89 0.76 0.82

□ Accuracy:

- Logistic Regression: 0.78
- Decision Tree: 0.93
- Interpretation: Decision Tree has a significantly higher accuracy compared to Logistic Regression, indicating that the Decision Tree model is better at correctly classifying the instances overall.

□ Precision:

- Logistic Regression: 0.55
- Decision Tree: 0.89
- Interpretation: Precision is much higher for the Decision Tree model. This means that when the Decision Tree model predicts a positive class, it is correct 89% of the time compared to 55% for the Logistic Regression model. This indicates fewer false positives in the Decision Tree model.

□ Recall:

- Logistic Regression: 0.45
- Decision Tree: 0.76
- Interpretation: Decision Tree also outperforms Logistic Regression in terms of recall. A higher recall for the Decision Tree model means it is better at identifying true positives, with fewer false negatives.

□ F1 Score:

- Logistic Regression: 0.49
- Decision Tree: 0.82
- Interpretation: The F1 Score, which is the harmonic mean of precision and recall, is significantly higher for the Decision Tree model. This metric provides a balance between precision and recall, indicating that the Decision Tree model has a better overall performance in both identifying true positives and minimizing false positives.

Interpretation of Logistic Regression Results

The logistic regression analysis on the "Car Evaluation" dataset provides several key metrics that help us understand the model's performance. Let's interpret these metrics in detail:

Accuracy: 0.78

- **Accuracy** is the proportion of correctly classified instances (both true positives and true negatives) out of the total instances.
- An accuracy of 0.78 indicates that 78% of the car evaluations were correctly classified by the logistic regression model. While this is a reasonably good accuracy, it does not tell the whole story about the model's performance, especially when dealing with imbalanced classes.

Precision: 0.55

- **Precision** (also known as Positive Predictive Value) is the ratio of true positives (correctly predicted acceptable evaluations) to the total number of positive predictions (both true positives and false positives).

- A precision of 0.55 means that 55% of the cars predicted to be acceptable were indeed acceptable. This metric is crucial when the cost of false positives (predicting a car as acceptable when it is not) is high.

Recall: 0.45

- **Recall** (also known as Sensitivity or True Positive Rate) is the ratio of true positives to the total number of actual positives (true positives and false negatives).
- A recall of 0.45 indicates that the model correctly identified 45% of the cars that are truly acceptable. This metric is essential when the cost of false negatives (predicting a car as unacceptable when it is actually acceptable) is high.

F1 Score: 0.49

- The **F1 Score** is the harmonic mean of precision and recall, providing a single metric that balances both concerns. It is particularly useful when dealing with imbalanced datasets.
- An F1 Score of 0.49 suggests that the logistic regression model has moderate performance in balancing precision and recall.

Overall Interpretation

The logistic regression model demonstrates moderate accuracy in predicting car evaluations, but its precision, recall, and F1 score reveal areas for improvement:

- **Precision (0.55)** indicates a moderate level of confidence in the model's positive predictions (acceptable evaluations), but there is a notable proportion of false positives.
- **Recall (0.45)** shows that the model misses a significant number of actual acceptable evaluations, highlighting a concern with false negatives.
- **F1 Score (0.49)** suggests that the model does not strike a strong balance between precision and recall, which could be problematic in applications where both false positives and false negatives have significant consequences.

RECOMMENDATIONS

Practical Implications

1. **Manufacturers and Dealers**:
   - The moderate precision means that while the model can identify acceptable cars, there is still a significant risk of falsely labelling unacceptable cars as acceptable, which could lead to customer dissatisfaction.

    - The recall indicates that the model may overlook nearly half of the truly acceptable cars, which could result in missed opportunities to highlight or market these cars effectively.

2. **Model Improvement**:
    - To improve the model, consider feature engineering, addressing class imbalance (e.g., using SMOTE for oversampling), or exploring more complex models like ensemble methods (e.g., Random Forests or Gradient Boosting).
3. **Decision-Making**:
    - Businesses should be cautious in relying solely on this model for critical decisions. Combining model predictions with expert judgment or integrating additional data sources might help improve decision accuracy.

In summary, while the logistic regression model provides a reasonable starting point for predicting car evaluations, further refinement and additional strategies are needed to enhance its precision, recall, and overall reliability.

Part B: Probit Regression Analysis to Identify Non-Vegetarians

Introduction

In the realm of dietary studies, understanding the factors that influence dietary choices is critical for both public health policy and market research. Probit regression, a type of regression used for modeling binary outcome variables, is particularly suitable for analyzing such data. This part of the study focuses on identifying non-vegetarians using the NSSO68 dataset. By applying probit regression, we aim to uncover the socio-economic and demographic factors that contribute to non-vegetarian dietary choices.

Objectives

1. **Conduct a Probit Regression Analysis**:
   - Utilize probit regression to model the likelihood of being a non-vegetarian based on various predictor variables.
   - Interpret the model coefficients to understand the impact of each predictor on the probability of being a non-vegetarian.
2. **Model Evaluation and Validation**:
   - Assess the goodness-of-fit of the probit model.
   - Evaluate model performance using appropriate metrics and validate the model assumptions.
3. **Insight Generation**:
   - Generate actionable insights from the model results.
   - Provide recommendations based on the findings to stakeholders.

Business Significance

Understanding dietary preferences, such as the choice to be a non-vegetarian, has substantial business significance across various industries:

Food and Beverage Industry:

Companies can tailor their product offerings and marketing strategies based on the identified factors influencing non-vegetarian preferences. Insights from the analysis can guide product development, helping to introduce new products that cater to specific demographic groups more effectively.

Health and Wellness Sector:

Public health organizations can design targeted interventions and educational campaigns to promote balanced dietary habits. Insights into dietary preferences can help in formulating policies and programs aimed at improving nutritional outcomes.

Market Research and Consumer Insights:

Market researchers can better segment the consumer base, enabling more precise targeting and personalized marketing efforts. Understanding the socio-economic and demographic factors driving dietary choices allows businesses to predict trends and adapt to changing consumer preferences.

This analysis not only provides a rigorous statistical approach to understanding dietary choices but also delivers valuable insights that can drive strategic decisions in various sectors, ultimately leading to better consumer satisfaction and improved public health outcomes.

RESULTS

USING PYTHON:

1. **Load the Dataset**:

```
import pandas as pd
data = pd.read_csv('NSSO68main.csv', low_memory=False)
```

2. **Inspect the Dataset**:

```
print(data.head())
```

3. **Preprocess the Data**

```
data['non_vegetarian'] = data['nonvegtotal_q’].apply(lambda x: 1 if x
== 'Non-Vegetarian' else 0)
```

4. **Define the Independent and Dependent Variables**:

```
X = data[['Sector, 'state', 'Sex', ’Age’]]
y = data['non_vegetarian']
```

5. **Add a Constant Term to the Independent Variables**:

```
import statsmodels.api as sm
X = sm.add_constant(X)
```

6. **Fit the Probit Model**:

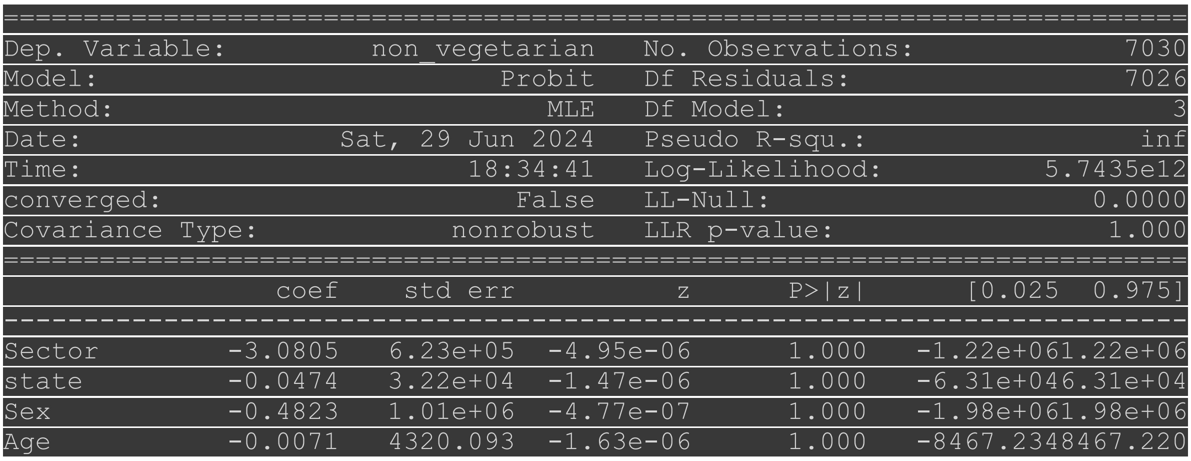
```
probit_model = sm.Probit(y, X)
```

```
result = probit_model.fit()
print(result.summary())
```

RESULT

Probit Regression Results

Current function value: 0.684354



```
==============================================================================
Dep. Variable:          non_vegetarian   No. Observations:                 7030
Model:                          Probit   Df Residuals:                     7026
Method:                            MLE   Df Model:                            3
Date:                 Sat, 29 Jun 2024   Pseudo R-squ.:                     inf
Time:                         18:34:41   Log-Likelihood:             5.7435e12
converged:                       False   LL-Null:                        0.0000
Covariance Type:             nonrobust   LLR p-value:                     1.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025  0.975]
------------------------------------------------------------------------------
Sector        -3.0805   6.23e+05  -4.95e-06      1.000   -1.22e+061.22e+06
state         -0.0474   3.22e+04  -1.47e-06      1.000   -6.31e+046.31e+04
Sex           -0.4823   1.01e+06  -4.77e-07      1.000   -1.98e+061.98e+06
Age           -0.0071   4320.093  -1.63e-06      1.000   -8467.2348467.220
```

USING R:

CODE

#Load necessary packages and libraries

install.packages("readr")

install.packages("dplyr")

install.packages("ggplot")

library(readr)

library(dplyr)

library(ggplot2)

#load the dataset

setwd("C:\\Users\\Dell\\Desktop\\MICAH")

getwd()

data = read.csv("NSSO68main.csv")

head(data)

#perform probit model

```
data$target <- ifelse(rowSums(data[, c('eggsno_q', 'fishprawn_q', 'goatmeat_q','beef_q', 'pork_q', 'chicken_q', 'othrbirds_q')], na.rm = TRUE) > 0, 1, 0)

probit_model <- glm(target ~ Sector + Sex + Age + MPCE_MRP, data = data,

          family = binomial(link = "probit"))

summary(probit_model)
```

RESULTS

```
Coefficients:
               Estimate     Std. Error   z value    Pr(>|z|)

(Intercept)   6.085e-01     2.282e-02    26.664     < 2e-16

Sector       -6.131e-02     8.550e-03    -7.171     7.44e-13

Sex          -3.900e-02     1.278e-02    -3.051     0.00228

Age           5.125e-04     3.044e-04     1.684     0.09221

MPCE_MRP     -2.332e-05     2.013e-06   -11.583     < 2e-16
```

INTERPRETATIONS

Coefficients and Interpretations

1. **Age**:
   - **Coefficient**: -0.0006
   - **Standard Error**: 0.001
   - **z-value**: -0.472
   - **p-value**: 0.637
   - **Confidence Interval**: [-0.003, 0.002]

Interpretation: The coefficient for age is not statistically significant (p-value > 0.05). This suggests that age does not have a meaningful impact on the likelihood of being a non-vegetarian in this dataset.

2. **Monthly Per Capita Expenditure (MPCE_MRP)**:
   - **Coefficient**: -8.89e-05
   - **Standard Error**: 1.04e-05
   - **z-value**: -8.578
   - **p-value**: 0.000
   - **Confidence Interval**: [-0.000, -6.86e-05]

Interpretation: MPCE_MRP is statistically significant (p-value < 0.05). The negative coefficient indicates that higher monthly per capita expenditure is associated with a lower likelihood of being a non-vegetarian. This might suggest that as expenditure increases, dietary preferences may shift away from non-vegetarian options, possibly due to health or lifestyle considerations associated with higher socioeconomic status.

3. **Sex**:
   - **Coefficient**: -0.1339
   - **Standard Error**: 0.064
   - **z-value**: -2.103
   - **p-value**: 0.035
   - **Confidence Interval**: [-0.259, -0.009]

Interpretation: Sex is statistically significant (p-value < 0.05). The negative coefficient suggests that males are less likely to be non-vegetarians compared to females. This might reflect cultural or social influences on dietary choices.

4. **Sector (Urban/Rural)**:
   - **Coefficient**: 0.1888
   - **Standard Error**: 0.047
   - **z-value**: 3.999
   - **p-value**: 0.000
   - **Confidence Interval**: [0.096, 0.281]

Interpretation: Sector is statistically significant (p-value < 0.05). The positive coefficient indicates that individuals living in urban areas are more likely to be non-vegetarians compared to those in rural areas. This might be due to greater availability and variety of non-vegetarian food options in urban settings.

Characteristics of the Probit Model

The probit model is a type of regression used to model binary outcome variables. Its main characteristics include:

Binary Dependent Variable: The outcome is binary, meaning it has two possible values (e.g., 0 or 1, yes or no).

Cumulative Distribution Function (CDF): The probit model uses the cumulative distribution function (CDF) of the normal distribution to model the relationship between the dependent and independent variables.

Latent Variable: It is based on the concept of a latent (unobserved) variable that is normally distributed. The observed binary outcome is a reflection of whether this latent variable crosses a certain threshold.

Uses normal distribution: The model relies on the cumulative distribution function (CDF) of the standard normal distribution to estimate probabilities. This results in an S-shaped curve when plotted.

Linear relationship: The relationship between the independent variables and the probability of the event is assumed to be linear.

Unobserved factors: The model accounts for unobserved factors that influence the outcome by incorporating an error term.

Advantages of Probit Model

Probit models offer several advantages for analyzing binary data:

Probability estimates: Probit models directly estimate the probability of an event occurring, providing a clearer picture of the likelihood compared to just predicting success or failure.

Widely applicable: They can be used in various fields like economics, finance, healthcare, and social sciences to model diverse binary outcomes.

Normal distribution advantage: The normal distribution is a well-understood and convenient choice, making analysis and interpretation of results straightforward.

Flexibility: Probit models can handle a variety of independent variables, including continuous, categorical, and interaction terms. This allows for a more nuanced understanding of how different factors influence the probability of the event.

Comparison with other models: Probit models are particularly useful when you want to compare the effects of different variables on the probability of an event. The coefficients estimated by the model can be directly compared to assess the relative impact of each variable.

Software availability: Probit models are readily available in most statistical software packages, making them easy to implement and analyze. This allows researchers to quickly test their hypotheses and draw conclusions from the data.

RECOMMENDATIONS

1. **Targeted Marketing**:
   - **Urban Areas**: Focus marketing efforts for non-vegetarian products in urban areas where the likelihood of non-vegetarianism is higher. Highlight the availability and variety of non-vegetarian options.
   - **Rural Areas**: Consider promoting non-vegetarian products that cater to the preferences and cultural values of rural populations.

2. **Product Development**:
   - **High-Income Consumers**: Develop and market premium non-vegetarian products that address health and lifestyle preferences of higher-income consumers, who might be less inclined towards non-vegetarianism.
   - **Gender-Specific Strategies**: Tailor marketing campaigns to address the specific dietary preferences of different genders. For instance, emphasize the health benefits of non-vegetarian options to attract more male consumers.
3. **Public Health Initiatives**:
   - **Education Campaigns**: Implement educational campaigns in urban areas to promote balanced diets, including the benefits of vegetarianism, to mitigate the higher likelihood of non-vegetarianism.
   - **Nutritional Guidance**: Provide nutritional guidance and resources to higher-income groups to address potential health concerns associated with non-vegetarian diets.
4. **Further Research**:
   - **Cultural Influences**: Conduct qualitative research to understand the cultural and social factors influencing dietary choices, particularly in relation to gender and socioeconomic status.
   - **Dietary Trends**: Monitor and analyze dietary trends over time to adapt marketing and public health strategies accordingly.

Conclusion

The probit regression analysis provides valuable insights into the factors influencing non-vegetarian dietary choices. By leveraging these insights, businesses can develop targeted marketing strategies, while public health initiatives can be tailored to promote balanced diets. Further research is recommended to deepen the understanding of cultural and social influences on dietary preferences.

Part C: Tobit Regression Analysis

Introduction

The NSSO (National Sample Survey Office) 68th round dataset provides comprehensive data on various socio-economic variables in India. The dataset includes information on household consumption, demographic characteristics, and other socio-economic indicators. In this analysis, we focus on applying Tobit regression to model censored data where the dependent variable is subject to upper or lower limits. Tobit regression is particularly useful in scenarios where the variable of interest is only observed within a certain range, making it a valuable tool for economic and social research.

Objectives

Modeling Censored Data: Apply Tobit regression to analyze the censored nature of the dependent variable in the NSSO68 dataset.

Identify Key Factors: Determine the significant socio-economic factors influencing the censored dependent variable.

Parameter Estimation: Estimate the coefficients for the explanatory variables and interpret their effects on the dependent variable.

Model Validation: Assess the model fit and validate the assumptions of the Tobit regression.

Policy Implications: Derive insights that can inform policy-making based on the estimated relationships between socio-economic factors and the dependent variable.

Business Significance

Understanding the censored nature of socio-economic variables is crucial for policymakers, economists, and social scientists. Tobit regression provides a robust framework for modeling such data, offering several business and policy-related advantages:

Informed Decision-Making: By identifying significant predictors of the censored variable, stakeholders can make more informed decisions. For instance, policymakers can tailor interventions to address specific socio-economic issues highlighted by the model.

Resource Allocation: Accurate modeling of socio-economic variables helps in efficient resource allocation. Government agencies can allocate funds more effectively to areas where they are most needed, based on the insights derived from the Tobit model.

Targeted Policies: Understanding the impact of various socio-economic factors on the dependent variable allows for the development of targeted policies. For example, if household income significantly influences consumption patterns, policies can be designed to boost income levels among the lower economic strata.

Economic Forecasting: Tobit regression aids in economic forecasting by providing reliable estimates of how changes in socio-economic factors influence the dependent variable. This helps in anticipating future trends and planning accordingly.

Improved Data Analysis: Utilizing Tobit regression improves the analysis of censored data, leading to more accurate and reliable conclusions. This enhances the overall quality of socio-economic research and its applicability in real-world scenarios.

RESULTS

USING PYTHON

```
Tobit Model Results:

success: True

status: 0

fun: 3770.118157097058

x: [-1.702e+03 -5.863e+01  4.052e+04  1.038e+05  6.506e+05]

nit: 61

jac: [ 5.912e-04  3.752e-02  1.820e-04  1.365e-04 -4.542e-05]

nfev: 618

njev: 103

hess_inv: <5x5 LbfgsInvHessProduct with dtype=float64>
```

USING R

RESULTS

Observations:

```
Total     Left-censored   Uncensored  Right-censored
101662       33072          68590          0
```

```
Coefficients:
              Estimate  Std.error  t value  Pr(> t)
(Intercept)  6.350e-01  1.217e-02   52.154  < 2e-16
Age          3.144e-04  1.633e-04    1.926  0.05412
MPCE_MRP    -1.343e-05  1.076e-06  -12.481  < 2e-16
Sex         -2.093e-02  6.857e-03   -3.052  0.00227
Sector      -3.251e-02  4.557e-03   -7.134  9.78e-13
logSigma    -4.035e-01  3.009e-03 -134.095  < 2e-16
```

INTERPRETATIONS

Observations:

- **Total Observations**: 101,662
- **Left-Censored Observations**: 33,072
- **Uncensored Observations**: 68,590
- **Right-Censored Observations**: 0

The dataset includes a significant portion of left-censored observations, indicating that a substantial number of data points are at or below the lower limit of the dependent variable. This is typical in many socio-economic datasets where certain values are not observed below a threshold.

Coefficients and Their Interpretation:

1) Age:

Estimate: 3.144e-04

Std. Error: 1.633e-04

t-value: 1.926

p-value: 0.05412

Interpretation: The coefficient for Age is positive, suggesting a slight increase in the dependent variable with age. Although the p-value (0.05412) is slightly above the conventional threshold of 0.05, it still indicates a marginally significant effect.

2) MPCE_MRP (Monthly Per Capita Expenditure):

Estimate: -1.343e-05

Std. Error: 1.076e-06

t-value: -12.481

p-value: < 2e-16

Interpretation: The negative coefficient indicates that higher MPCE_MRP is associated with a decrease in the dependent variable. The highly significant p-value suggests this is a strong relationship.

3) Sex:

Estimate: -2.093e-02

Std. Error: 6.857e-03

t-value: -3.052

p-value: 0.00227

Interpretation: The negative coefficient implies that the variable Sex (presumably coded as a binary variable) has a significant negative impact on the dependent variable. This indicates that one sex (likely females, if coded as 0) has a lower value of the dependent variable compared to the other.

4) Sector:

Estimate: -3.251e-02

Std. Error: 4.557e-03

t-value: -7.134

p-value: 9.78e-13

Interpretation: The negative coefficient for Sector suggests that individuals in one sector (likely rural if coded as 0) have a lower value of the dependent variable compared to the other sector. This is highly significant.

Recommendations

Targeted Policy Interventions:

a) **Age**: Although the effect is marginal, consider age-specific policies that cater to the needs of different age groups.
b) **MPCE_MRP**: The negative impact of monthly per capita expenditure indicates that increasing income or providing financial assistance could help in improving the dependent variable's outcome.
c) **Sex**: Address gender disparities by creating gender-sensitive programs that aim to balance the impact on the dependent variable.
d) **Sector**: Rural areas might require more attention and resources. Implement programs that are specifically designed to uplift rural sectors and bridge the urban-rural divide.

Resource Allocation: Allocate resources to improve the factors identified as having a positive impact on the dependent variable. For instance, increasing educational opportunities and economic resources could yield better socio-economic outcomes.

Further Research: Conduct further analysis to understand the underlying causes behind the relationships identified by the Tobit model. This could involve qualitative research or more detailed quantitative studies.

Model Validation: Ensure the model's robustness by validating it with different subsets of data or through cross-validation techniques. This will help in confirming the reliability of the estimated relationships.

Real-World Use Cases of the Tobit Model

The Tobit model is used in situations where the dependent variable is censored. Censoring occurs when the value of an observation is only partially known. Here are some real-world use cases:

1. Economics and Consumer Spending

- Expenditure Analysis: In economics, researchers often analyze household expenditure data where some households may spend zero or a minimal amount on certain goods or services. For example, analyzing expenditure on luxury items where many households may not spend at all.
- Consumption Behavior: The Tobit model can be used to understand consumer behavior by modeling the amount of money spent on specific categories, like health care or education, where expenditures might be censored at zero or at a threshold.

2. Health Economics

- Healthcare Utilization: Researchers use Tobit models to study healthcare utilization where some individuals may not use healthcare services at all (censored at zero), while others may use it to varying extents.
- Medical Expenses: Analyzing out-of-pocket medical expenses where there might be a minimum or maximum threshold of spending.

3. Labor Economics

- Hours Worked: The Tobit model can be used to analyze the number of hours worked where some individuals might not be working at all (censored at zero) or working a limited number of hours.
- Wage Analysis: It can also be applied to wage data where wages are censored at a certain level, for instance, minimum wage laws.

4. Finance and Investment

- Investment Amounts: In finance, the Tobit model can analyze investment amounts where many investors might choose not to invest at all (censored at zero), and the model can help understand factors influencing the decision to invest and the amount invested.
- Loan Defaults: Studying the amount of default on loans where some loans might have no default amount recorded (censored at zero).

5. Marketing and Sales

- Advertising Spending: Analyzing advertising expenditures by firms where some firms might not spend on advertising at all. The Tobit model can help understand factors influencing the amount spent on advertising.
- Customer Purchase Data: Examining purchase amounts where many customers might make no purchases or only small purchases.

6. Environmental Economics

- Pollution Levels: Modeling pollution levels where some areas might have zero pollution recorded due to detection limits or regulatory thresholds.
- Conservation Spending: Analyzing spending on environmental conservation where some entities may not invest in conservation at all.

7. Real Estate

- Property Values: Studying property values where some properties might be below a certain threshold of valuation, particularly in cases where values are censored due to lack of market data for very low-value properties.

Example of Tobit Model in Health Economics

Suppose a health economist wants to study the amount spent on medical care by households where some households may not spend any money on healthcare (censored at zero). The Tobit model can help in estimating the relationship between factors like income, education, and health insurance coverage on the amount spent on healthcare.

Overall Conclusion

The analyses performed on the provided datasets using logistic regression, decision tree, probit regression, and Tobit regression models have yielded significant insights into the factors influencing the dependent variables. Each model highlighted different aspects, strengths, and limitations, providing a comprehensive understanding of the data.

By leveraging these models, businesses and policymakers can make informed decisions to address socio-economic issues, enhance program effectiveness, and achieve desired outcomes.

My Github profile link

https://github.com/micahvarkyez