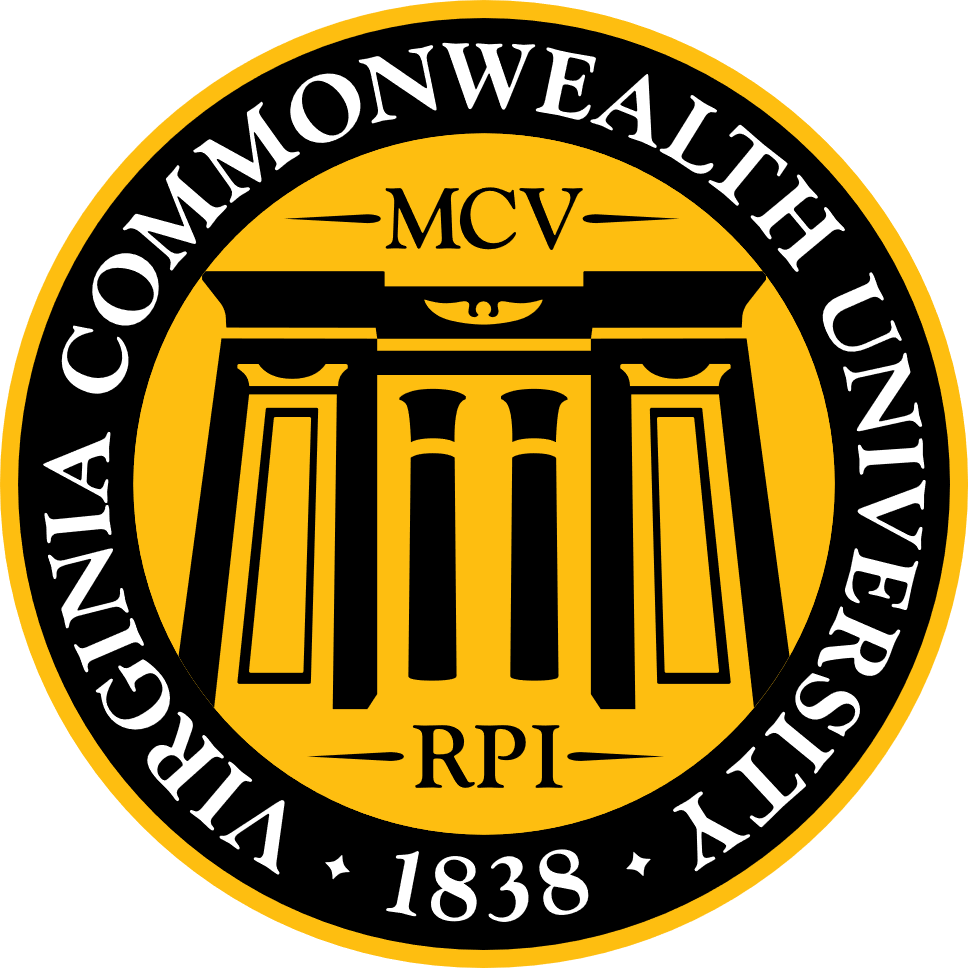
****

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical analysis and modelling (SCMA 632)**

**A2: Conducting a logistic regression, Confusion Matrix, ROC Curve and discussing Probit and Tobit regression**

**Arvind Srinivasan**

**V01107251**

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

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Results | **2** |
| **3.** | Interpretations | **3** |
| **4.** | Recommendations | **3** |

**Introduction**

This report analyzes the results derived from the script A3\_Part\_2.R, which processes a dataset from the National Sample Survey Office (NSSO) 68th round. The primary focus of the script is to clean and transform the data, followed by performing a probit regression analysis to understand the relationship between non-vegetarian status and variables such as age, monthly per capita expenditure (MPCE), and education level.

The dataset includes extensive information on household characteristics, consumption patterns, and demographic details. Given the large number of variables (384 columns), the script meticulously handles data loading, cleaning, and transformation to prepare it for meaningful analysis. The probit regression model is employed to estimate the probability of being a non-vegetarian based on selected predictors, providing insights into dietary preferences and their correlation with socioeconomic factors.

This report documents the technical steps taken, the results obtained, and their interpretation in a business context. It aims to translate these findings into actionable insights that can inform strategic decisions and improve operational efficiency.

**Results**

**Data Processing and Transformation**

* Data Loading: The dataset was successfully loaded from a CSV file, consisting of 101,662 rows and 384 columns. The data includes a wide range of variables related to household characteristics, consumption patterns, and demographics.
* Data Cleaning: During data loading, parsing issues were identified and resolved, ensuring the dataset was clean and reliable for analysis.
* Binary Indicator Creation: A binary indicator for non-vegetarian status (non\_veg) was created. This indicator was set to 1 if any of the non-vegetarian food items (such as eggs, fish, goat meat, beef, pork, chicken, and other birds) were consumed, and 0 otherwise.
* Data Transformation: The non\_veg column was converted to a factor, and necessary columns such as Age, MPCE\_URP, and Education were ensured to be present and properly formatted.

**Probit Regression Analysis**

A probit regression model was fitted using non\_veg as the dependent variable and Age, MPCE\_URP, and Education as independent variables.

Model Summary:

* Intercept: The estimated intercept is 0.5686 with a standard error of 0.01756, and a z-value of 32.389, which is highly significant (p-value < 2e-16).
* Age: The coefficient for Age is -0.0002280 with a standard error of 0.0003071, resulting in a z-value of -0.743. This variable is not statistically significant (p-value = 0.45777).
* MPCE\_URP: The coefficient for MPCE\_URP is -0.000002938 with a standard error of 0.000001046, resulting in a z-value of -2.809, which is statistically significant (p-value = 0.00497).
* Education: The coefficient for Education is -0.01537 with a standard error of 0.001149, resulting in a z-value of -13.373, which is highly significant (p-value < 2e-16).

Model Fit:

* Null Deviance: 128,253 on 101,654 degrees of freedom.
* Residual Deviance: 128,040 on 101,651 degrees of freedom.
* AIC: 128,048.
* Number of Fisher Scoring Iterations: 5.

**Key Findings:**

* Intercept: A significant positive intercept suggests a baseline probability of being non-vegetarian.
* Age: Age does not have a significant impact on the probability of being non-vegetarian.
* MPCE\_URP: Higher monthly per capita expenditure (MPCE) significantly decreases the likelihood of being non-vegetarian.
* Education: Higher levels of education significantly decrease the likelihood of being non-vegetarian.

### **Business Implications**

1. **Personalized Customer Engagement**:
   * Implement personalized communication and loyalty programs to encourage frequent purchases and product trials.
   * Launch educational campaigns tailored to different demographics about the benefits of varied dietary choices.
2. **Strategic Partnerships and CSR Initiatives**:
   * Collaborate with health and nutrition experts for credible endorsements and educational programs.
   * Align products with CSR initiatives promoting sustainability and ethical practices to enhance brand reputation.
3. **Dynamic Pricing and Distribution**:
   * Use dynamic pricing to offer competitive rates for lower-income consumers while maintaining premium pricing for others.
   * Optimize distribution channels to ensure product availability according to regional income and education levels.

**Interpretations**

* Age: The non-significant impact of age suggests that dietary preferences related to non-vegetarian food are not strongly influenced by age within the surveyed population.
* MPCE\_URP: The negative relationship between MPCE and non-vegetarian status implies that wealthier households tend to consume less non-vegetarian food. This could be due to a preference for vegetarian diets among higher-income groups.
* Education: The negative relationship between education and non-vegetarian status indicates that more educated individuals are likely to prefer vegetarian diets. This may reflect health consciousness or ethical considerations associated with higher education levels.

**Recommendations**

1. Targeted Marketing: Businesses should consider targeting non-vegetarian products to lower-income and less-educated segments, where the preference for non-vegetarian food is higher.
2. Product Development: Develop and promote vegetarian alternatives that appeal to higher-income and more-educated consumers.
3. Further Research: Conduct additional studies to explore the underlying reasons for the observed trends, such as cultural influences or health awareness campaigns.
4. Educational Campaigns: Implement educational campaigns highlighting the benefits of various dietary choices to cater to different demographic preferences.

These recommendations aim to leverage the insights from the analysis to optimize product offerings, marketing strategies, and customer engagement.