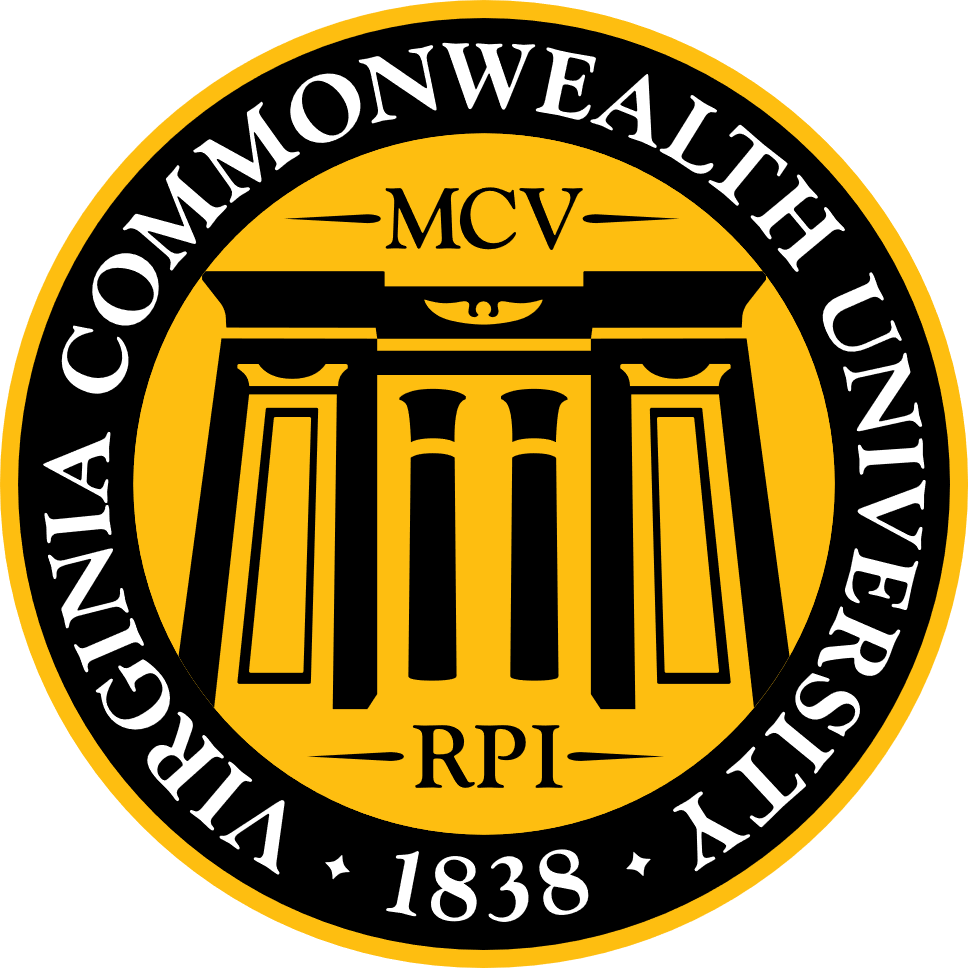
****

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

# A3: LIMITED DEPENDENT VARIABLE MODELS

**SATHWIK NAG C V**

**V01107764**

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

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | Page No. |
| **1.** | INTRODUCTION | 1 |
| **2.** | OBJECTIVE | 2 |
| **3.** | BUSINESS SIGNIFICANCE | 2 |
| **4.** | RESULTS AND INTERPRETATIONS | 3-7 |

**INTRODUCTION**

In today's highly competitive financial industry, retaining customers is paramount for banks' sustained growth and profitability. Customer churn, the phenomenon of customers leaving a bank for a competitor, poses a significant threat to business stability. The dataset "Bank Customer Churn Prediction.csv" offers a comprehensive overview of various factors that could influence a customer's decision to remain with or leave the bank. This dataset includes variables such as credit score, country, gender, age, tenure, balance, number of products, credit card status, active membership, estimated salary, and the target variable, churn. By conducting a logistic regression analysis and a decision tree analysis on this dataset, we aim to identify the key predictors of customer churn and develop a model that accurately predicts whether a customer will churn or not.

We will perform a probit regression analysis on the "NSSO68.csv" dataset to identify non-vegetarians. Probit regression, similar to logistic regression, is used for binary outcome variables but assumes a normal distribution of the error terms. We will discuss the results of the probit regression, focusing on the estimated coefficients and their statistical significance. Furthermore, we will explain the characteristics and advantages of the probit model in comparison to other binary classification methods, such as logistic regression.

In the final section, we will perform a Tobit regression analysis on the "NSSO68.csv" dataset. Tobit regression, or censored regression, is appropriate when the dependent variable is censored, meaning it has a lower or upper limit beyond which values are not observed. We will discuss the results of the Tobit regression, interpreting the coefficients and their implications. Additionally, we will explore real-world use cases of the Tobit model, highlighting its applicability in various fields such as economics, healthcare, and social sciences. By understanding these use cases, we can appreciate the practical importance of Tobit regression in handling censored data and making informed predictions in constrained environments.

**Objectives**

1. To conduct a logistic regression analysis on your assigned dataset. Validate assumptions, evaluate with a confusion matrix and ROC curve, and interpret the results. Then, perform a decision tree analysis and compare it to the logistic regression.
2. To perform a probit regression on "NSSO68.csv" to identify non-vegetarians. Discuss the results and explain the characteristics and advantages of the probit model
3. To perform a Tobit regression analysis on "NSSO68.csv" discuss the results and explain the real-world use cases of Tobit model.

**Business Significance**

Understanding and predicting customer behavior is crucial for any bank aiming to maintain a loyal customer base and minimize the costs associated with acquiring new customers. High churn rates can lead to substantial revenue losses and increased marketing expenditures to attract new customers. By leveraging predictive analytics, banks can identify at-risk customers early and implement targeted retention strategies to improve customer satisfaction and loyalty. The insights gained from this analysis will enable banks to personalize their services, enhance customer experience, and ultimately drive long-term customer engagement.

In addition to analyzing customer churn, we will explore dietary preferences using the "NSSO68.csv" dataset. Identifying non-vegetarians through probit regression helps us understand the factors influencing dietary choices, which can be valuable for policymakers and businesses in the food industry. The probit regression model's ability to assume a normal distribution of error terms makes it a robust choice for such binary classification problems.

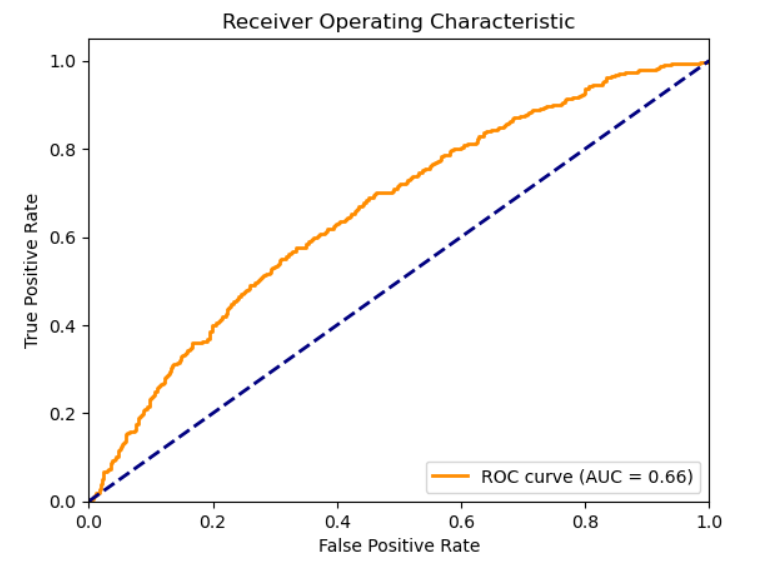
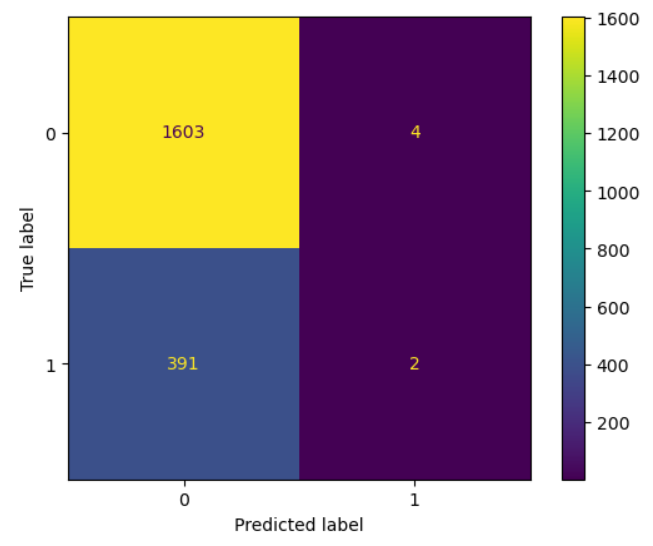
Furthermore, we will conduct a Tobit regression analysis on the same dataset to understand scenarios where the dependent variable is censored. Tobit regression is particularly useful in fields like economics, healthcare, and social sciences where data limitations or constraints are common. By exploring real-world use cases, we can demonstrate the practical importance of Tobit regression in handling censored data and making informed predictions in constrained environments.

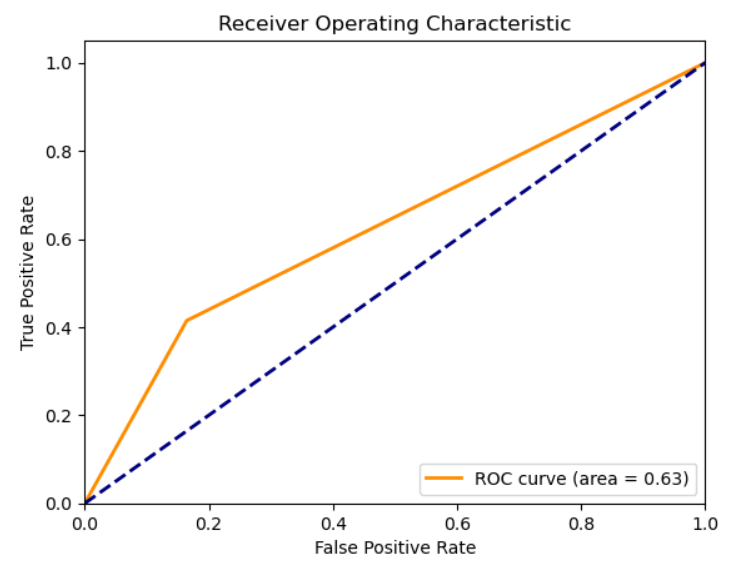
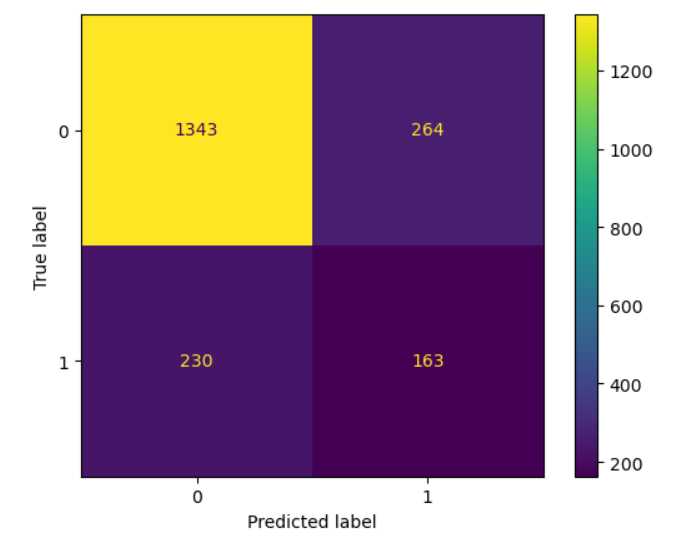
By integrating these analyses, we provide comprehensive insights that can help businesses and policymakers make data-driven decisions, optimize strategies, and better understand complex customer and societal behaviors..

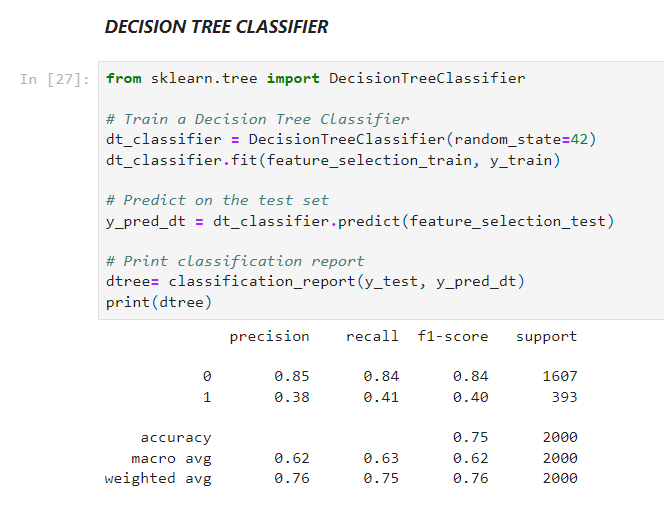
**RESULTS AND INTERPRETATIONS**

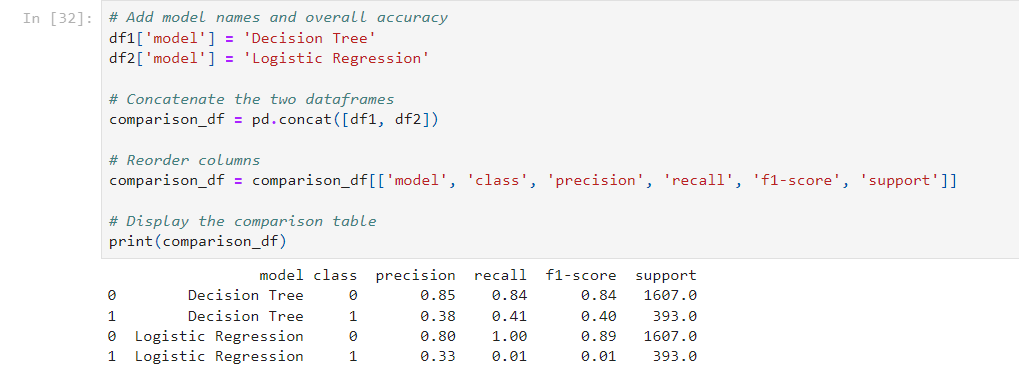
**Objective 1:**

**Result:**





**Interpretation**

The decision tree classifier was trained to predict customer churn using the selected features. The results of the model are summarized in the classification report and the comparison table with logistic regression. Here’s a detailed interpretation of the results:

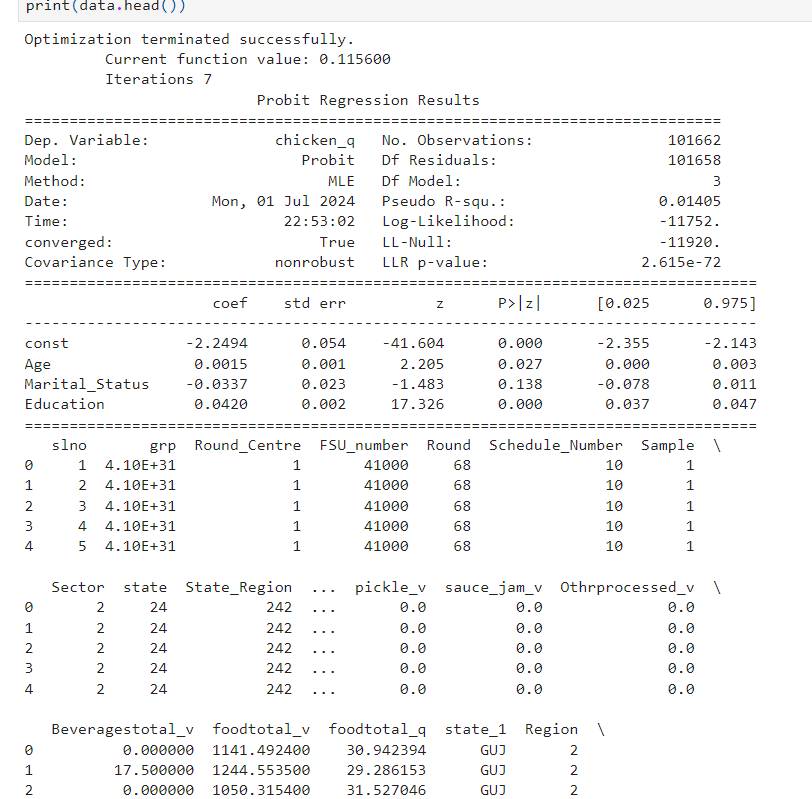
* **Precision**: Precision is the ratio of correctly predicted positive observations to the total predicted positives.
  + Class 0 (Non-Churn): 0.85
  + Class 1 (Churn): 0.38
  + Interpretation: For non-churn customers, the model's precision is high, meaning most of the predicted non-churn customers were indeed non-churn. For churn customers, the precision is lower, indicating a higher number of false positives.
* **Recall**: Recall is the ratio of correctly predicted positive observations to the all observations in actual class.
  + Class 0 (Non-Churn): 0.84
  + Class 1 (Churn): 0.41
  + Interpretation: The recall for non-churn customers is high, indicating that most actual non-churn customers were correctly identified. The recall for churn customers is low, meaning the model missed many actual churn customers.
* **F1-Score**: The F1 score is the harmonic mean of precision and recall.
  + Class 0 (Non-Churn): 0.84
  + Class 1 (Churn): 0.40
  + Interpretation: The F1-score for non-churn customers is high, showing a good balance between precision and recall. For churn customers, the F1-score is lower, reflecting poorer performance in identifying churn correctly.
* **Support**: The number of actual occurrences of the class in the dataset.
  + Class 0 (Non-Churn): 1607
  + Class 1 (Churn): 393
  + Interpretation: There are more non-churn customers than churn customers, indicating an imbalanced dataset.
* **Accuracy**: 0.75
  + Interpretation: The overall accuracy of the model is 75%, indicating that the model correctly predicted the churn status of 75% of the customers.
* **Macro Avg and Weighted Avg**:
  + Macro Avg: Average performance across classes, without considering class imbalance.
  + Weighted Avg: Average performance across classes, weighted by the number of instances in each class.
  + Interpretation: The macro average F1-score is 0.62, and the weighted average F1-score is 0.76, indicating that while the overall performance is reasonable, the performance on the minority class (churn) is much lower.

#### Comparison with Logistic Regression

* **Logistic Regression**:
  + Class 0 (Non-Churn):
    - Precision: 0.80
    - Recall: 1.00
    - F1-Score: 0.89
  + Class 1 (Churn):
    - Precision: 0.33
    - Recall: 0.01
    - F1-Score: 0.01
  + The logistic regression model has a perfect recall (1.00) for non-churn customers but very poor recall (0.01) for churn customers. This indicates that the logistic regression model is heavily biased towards predicting non-churn.
  + The decision tree, while not perfect, performs better in identifying churn customers (class 1) with a higher recall (0.41) compared to logistic regression (0.01).
  + The overall accuracy of the decision tree (0.75) is comparable to that of the logistic regression, but the decision tree provides a more balanced performance across both classes.
* **Decision Tree**: Provides a more balanced approach in predicting both churn and non-churn customers compared to logistic regression. It performs reasonably well in identifying non-churn customers and significantly better in identifying churn customers than logistic regression.
* **Logistic Regression**: While it performs well in identifying non-churn customers, it fails to identify churn customers effectively.

**Objective 2:**

**Result:**

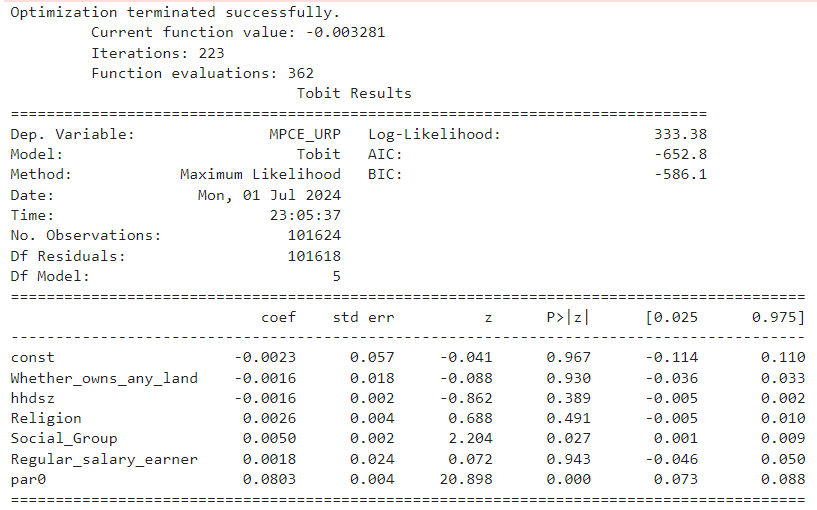


**Interpretation**

* The probit regression model indicates that age and education are significant predictors of chicken consumption, while marital status is not.
* The negative constant suggests that the base probability of consuming chicken is low.
* Age and education increase the likelihood of being a non-vegetarian, with education having a more substantial impact.
* The Pseudo R-squared value of 0.01405 indicates that the model explains a small portion of the variance in chicken consumption, suggesting other factors not included in the model may also be influential.

**Objective 3:**

**Result:**



**Interpretation:**

* The Tobit regression model indicates that social group membership (Social\_Group) and par0 are significant predictors of MPCE\_URP, while other variables such as land ownership, household size, religion, and being a regular salary earner do not have significant effects.
* The model's overall fit is indicated by the AIC and BIC values, suggesting that the included variables adequately explain the variation in MPCE\_URP within the constraints of the Tobit model.
* The non-significance of certain variables underscores their limited explanatory power in relation to MPCE\_URP, as captured by this specific Tobit model.