

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

**A3a : Data Analysis - Logistic vs Tree**

**Azim Ziyan A V01102412**

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

# TABLE OF CONTENTS

1. Introduction 3
2. Objectives 3

3. Business Significance ------------------------------------------------ 4

4. Results and Interpretations ------------------------------------------------ 4

# INTRODUCTION

The aim of this analysis is to build a logistic regression model using R and python to predict the loan status (approved or denied) for applicants based on a set of features available in the dataset. The dataset contains information about applicants, including their income, loan amount, credit history, and other relevant factors. The target variable, Loan\_Status, indicates whether a loan was approved (Y) or denied (N). Logistic regression is a statistical method used for binary classification that models the probability of a binary outcome based on one or more predictor variables. In the context of predicting loan status, logistic regression in R can help determine the likelihood of a loan application being approved (`Loan\_Status`). By analyzing various factors such as applicant income, loan amount, credit history, and other relevant variables, logistic regression quantifies the relationship between these predictors and the probability of loan approval. The R programming language, with its rich set of libraries like `dplyr` for data manipulation and `glm` for model fitting, offers a robust platform for performing logistic regression. This approach not only provides a probabilistic framework for decision-making but also allows for the evaluation of model performance using metrics such as accuracy, precision, recall, and the ROC-AUC curve, thereby aiding in the refinement of lending strategies and risk assessment.

# OBJECTIVES

a) 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

# BUSINESS SIGNIFICANCE

Customer Segmentation:

By predicting loan eligibility (Loan\_Status), the bank can effectively segment customers into those likely to be approved for loans and those who might require further scrutiny. This segmentation helps in targeting marketing efforts and improving customer service.

* Risk Management:

Identifying patterns in features that lead to loan approval or rejection helps the bank refine its risk assessment models, thereby reducing the likelihood of defaults and improving the overall quality of the loan portfolio.

* Resource Allocation:

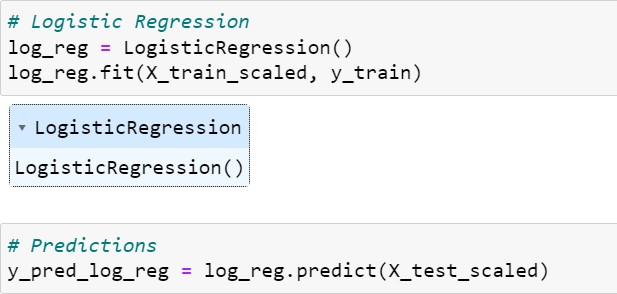
Predictive modeling allows the bank to streamline its loan approval process, saving time and resources. Automating parts of the process based on model predictions can lead to faster decision-making and better resource management.

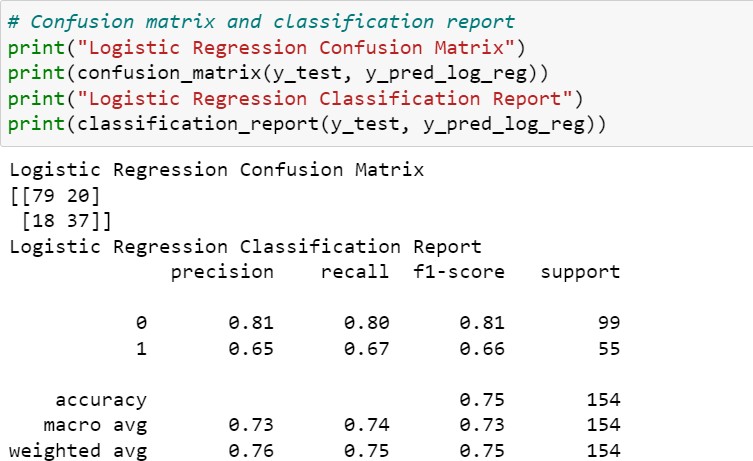
* Strategic Decision Making:

The insights from the models can inform strategic decisions regarding loan product offerings, interest rates, and terms, aligning them more closely with customer profiles and risk levels..

# RESULTS AND INTERPRETATIONS

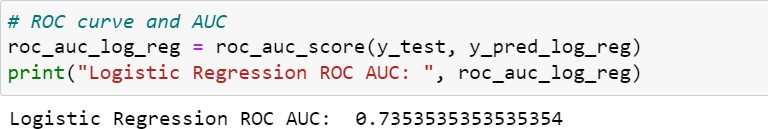
* Python

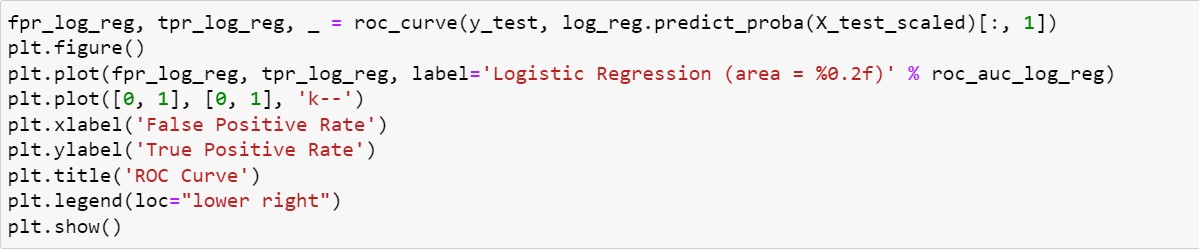


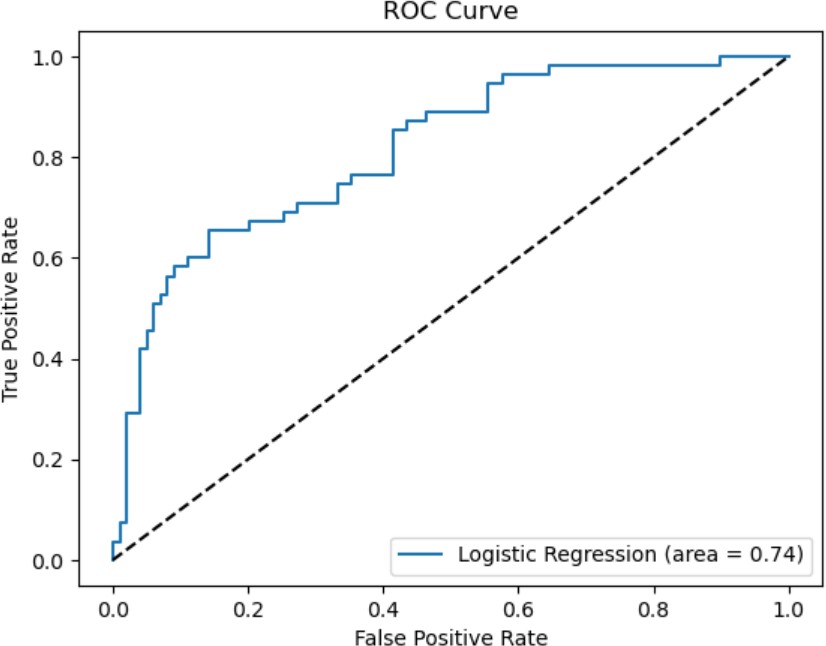


Interpretation

The confusion matrix for the logistic regression model reveals that the model correctly predicted 79 instances of rejection (true negatives) and 37 instances of approval (true positives). However, it incorrectly predicted diabetes in 20 cases where there was none (false positives) and failed to predict loan status in 18 cases where it was present (false negatives). These results indicate the model's performance in distinguishing between the two classes, showing a reasonable ability to correctly identify both approval and rejection of loan .





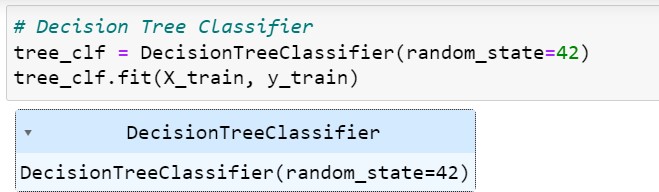


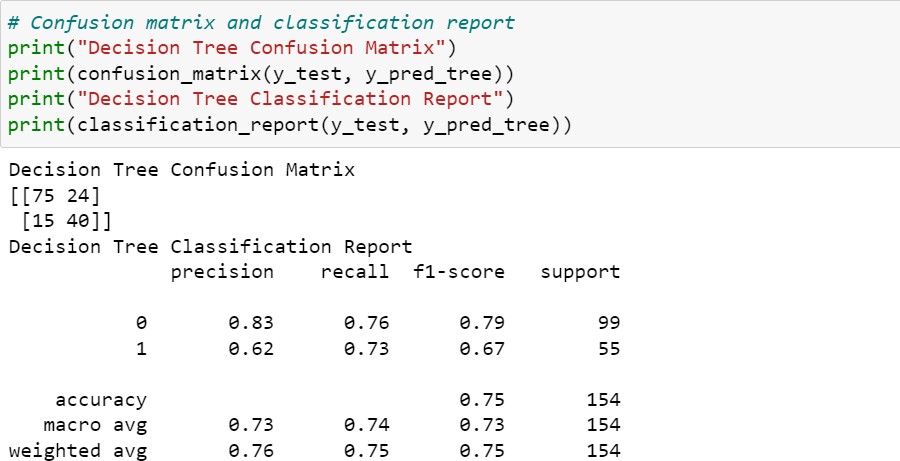
Interpretation

This code calculates the Area Under the Curve (AUC) for the Receiver Operating Characteristic (ROC) curve of the logistic regression model. The roc\_auc\_score function from the sklearn.metrics module is used to compute this metric, which evaluates the model's ability to distinguish between the positive and negative classes. The y\_test variable contains the true labels, while y\_pred\_log\_reg contains the predicted labels from the logistic regression model.

The ROC AUC value of approximately 0.74 indicates the overall performance of the logistic regression model in distinguishing between loan approval and rejection. An AUC value of

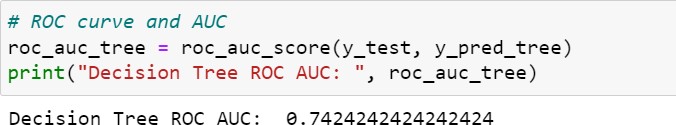
0.74 suggests that the model has a good but not perfect ability to discriminate between the two classes. In general, an AUC value closer to 1.0 indicates excellent model performance, while a value closer to 0.5 indicates performance no better than random chance.

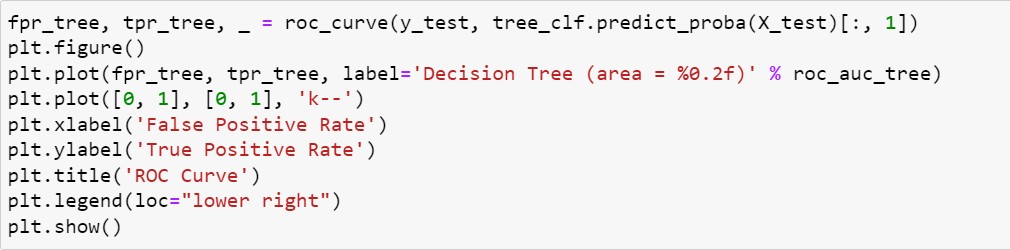


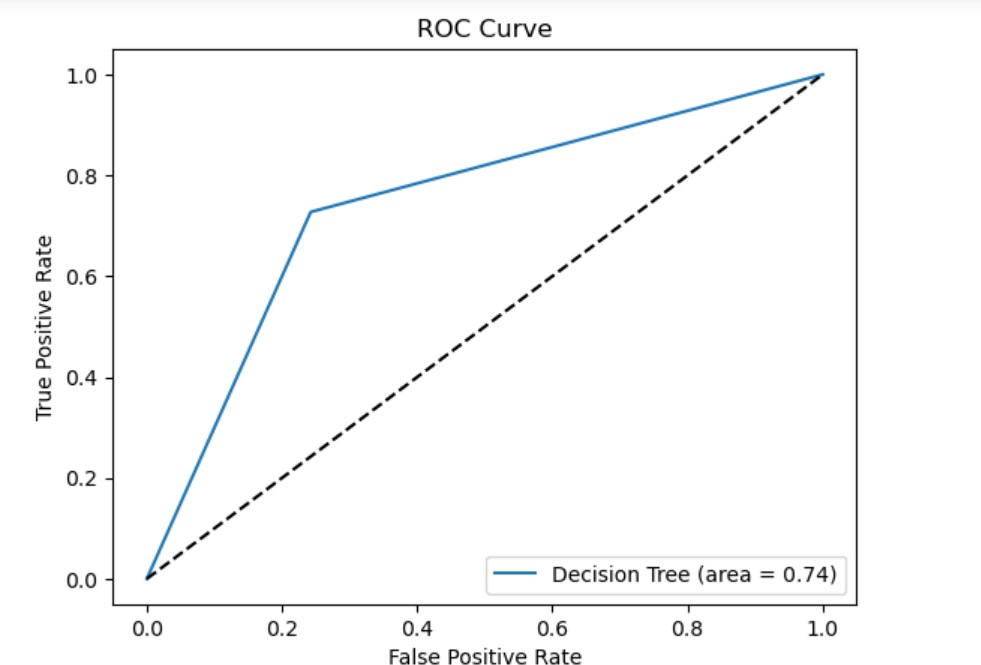


Interpretation

The confusion matrix for the decision tree model indicates that the model correctly predicted 75 instances of not approved (true negatives) and 40 instances of approval (true positives). However, it incorrectly predicted loan status in 24 cases where there was none (false positives) and failed to predict loan status in 15 cases where it was present (false negatives). These results show that while the decision tree model is effective at identifying both approval and rejection of loan status

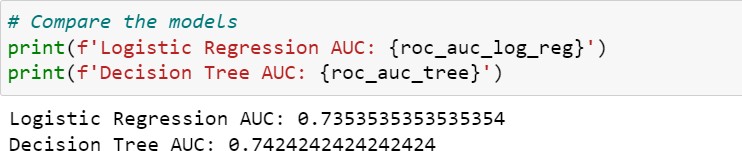






Interpretation

The decision tree model demonstrates a reasonable level of performance with an AUC of 0.74, indicating it is fairly effective at distinguishing between patients with and without approval of loan. This metric complements the previously discussed confusion matrix and classification report, providing a more comprehensive understanding of the model's predictive capabilities.



Interpretation

Both models exhibit similar performance in distinguishing between approval and rejection of loan cases, with AUC values slightly above 0.73, indicating good but not perfect discriminatory power.

Logistic Regression

The AUC of approximately 0.73 suggests that the logistic regression model has a good ability to differentiate between the two classes.

Decision Tree

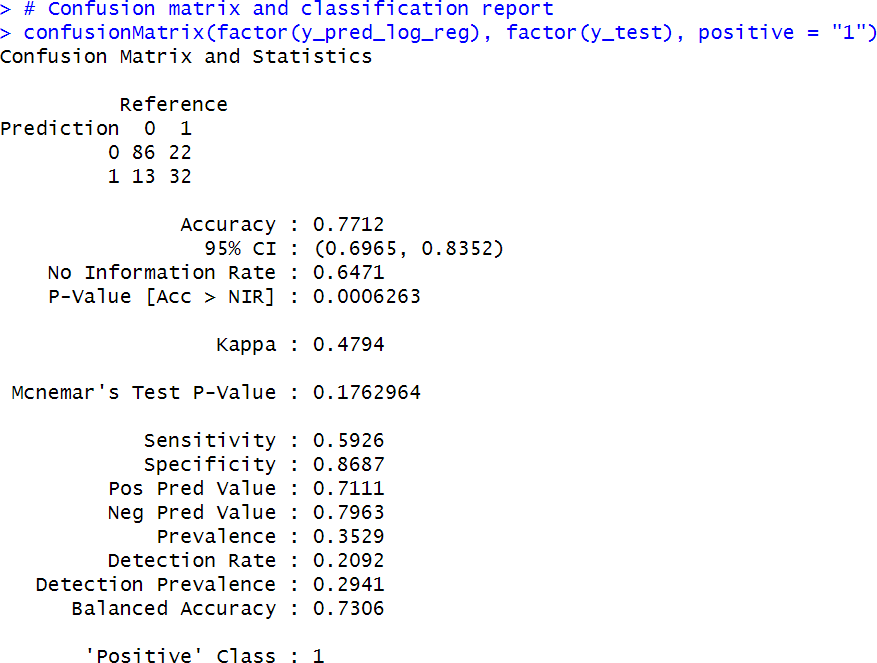
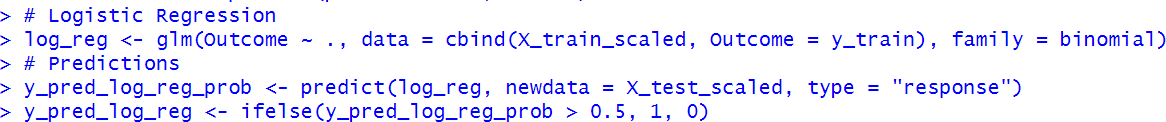
The decision tree model shows a slightly higher AUC of approximately 0.74, indicating marginally better performance compared to logistic regression.

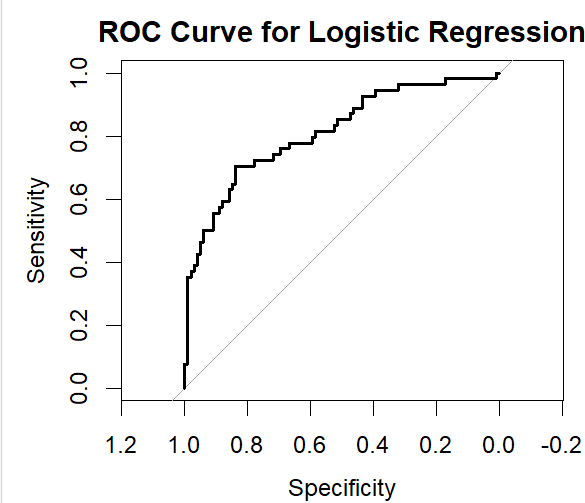
Overall Comparison

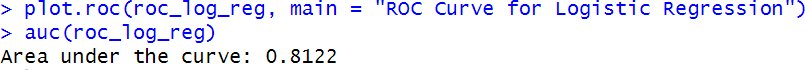
The marginal difference in AUC values (0.74 for the decision tree vs. 0.73 for logistic regression) suggests that both models perform similarly in terms of overall predictive accuracy.

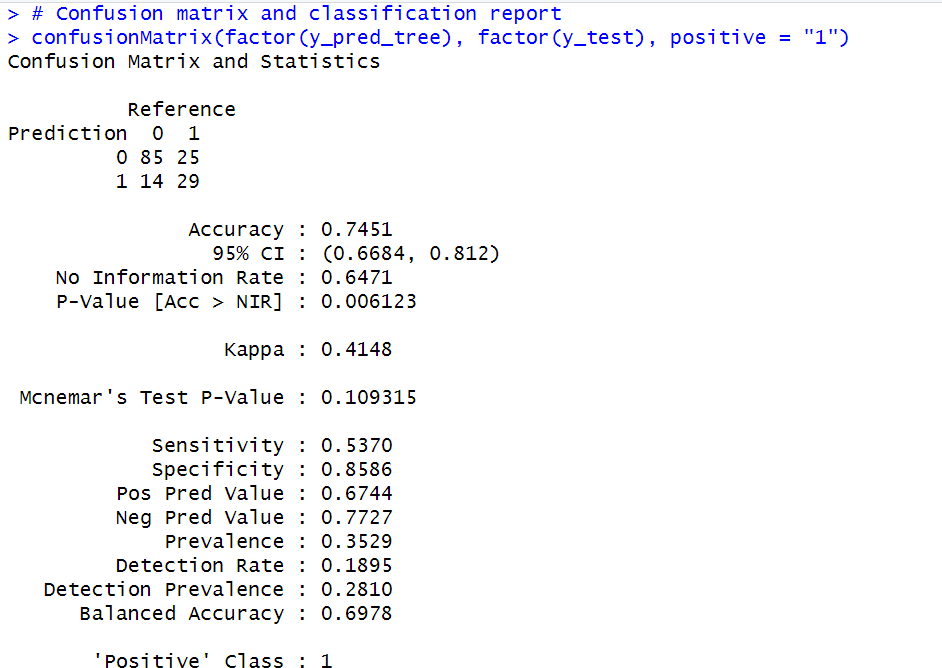
The choice between the two models may depend on other factors such as the importance of model interpretability, the ability to capture non-linear relationships, and the specific application context. For instance, if interpretability and understanding the influence of individual predictors are crucial, logistic regression might be preferred. Conversely, if the goal is to capture complex interactions and provide a straightforward decision-making process, a decision tree could be more suitable.

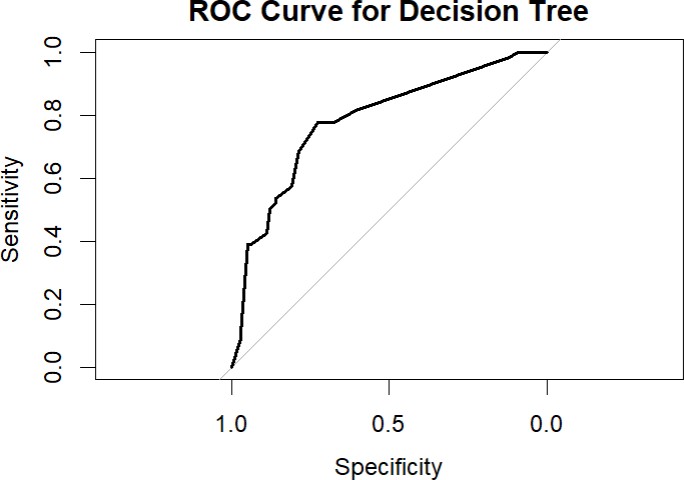
* R

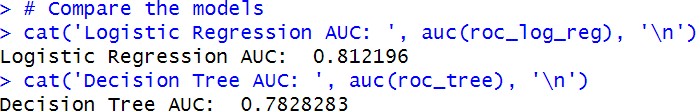
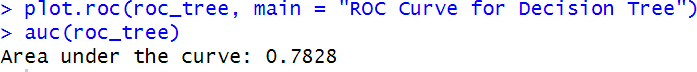












Interpretation

The provided AUC values for the logistic regression and decision tree models are 0.812196 and 0.7828283 respectively.

Logistic Regression

* The AUC of approximately 0.81 indicates that the logistic regression model has a strong ability to differentiate between approval and rejection of loan
* This higher AUC value suggests that the logistic regression model is effective in predicting the likelihood of loan approval, making it a reliable choice for classification tasks where model interpretability and understanding of individual predictors are important.
* Logistic regression provides clear insights into the relationship between each predictor variable and the outcome, which can be valuable for healthcare providers in understanding risk factors for loan approval.

Decision Tree

* The AUC of approximately 0.78 for the decision tree model indicates good performance but slightly lower than that of the logistic regression model.
* Decision trees are useful for capturing complex, non-linear relationships between variables and offer a visual representation of the decision-making process. This can be advantageous for intuitive understanding and explanation.
* Despite its slightly lower AUC, the decision tree model still demonstrates reasonable accuracy and may be preferred in scenarios where interpretability and capturing interactions between variables are crucial.

Overall Comparison

* The logistic regression model outperforms the decision tree model with a higher AUC (0.81 vs. 0.78), suggesting that it has a better overall ability to correctly classify approval and rejection of loan
* The logistic regression model’s higher AUC reflects its effectiveness in prediction and its robustness in handling linear relationships between the predictors and the outcome.
* On the other hand, the decision tree model, while slightly less accurate in terms of AUC, offers advantages in terms of visual interpretability and the ability to model complex relationships.