

Assignment 4

D	D	M	M	Y	Y

Q1] In the context of a financial institution, you are tasked with predicting loan default risk. Build a classification and regression tree (CART) model using historical loan data. Explain how you would split the data, choose the features and evaluate the model's performance using the confusion matrix and ROC curve. Discuss any ethical considerations related to using such a model.

Objective: Develop CART model for loan default prediction and evaluate it using various performance metrics while considering ethical implications.

⇒ Building a CART Model for loan Default prediction

Data splitting

The first step is to divide the historical loan data in 2 sets

Training set ⇒ This is used to build the cart model, it should be representative of overall population

Testing set ⇒ This dataset is used to evaluate the model's performance on unseen data

The common approach is to split the data in 80 : 20 proportion

Feature selection

Feature selection is crucial for building an effective CART model. Relevant features can significantly

improve model performance and interpretability, some key features for loan default prediction include

- Demographic information
- Financial information
- Behavioral information
- Economic indicators

Feature selection can be done using techniques like

- Correlation analysis
- Information gain
- Recursive Feature Elimination

Building the CART model

- Once the data is split and features are selected, a CART model can be constructed. The CART algorithm recursively partitions the data into subsets based on feature values.
- The splitting criterion is typically the Gini impurity or entropy. The process continues until a stopping criterion is met, such as reaching a maximum depth or minimum number of observations in a node.

Model Evaluation

- Confusion matrix: This matrix summarizes the model's predictions on the testing set. It helps to calculate accuracy, precision, recall and F1 score.

→ ROC : The receiver operating characteristic (ROC) curve plots the true positive rate against the false positive rate at different classification thresholds. The Area Under Curve (AUC) provides an overall measure of model performance.

Ethical consideration

- Fairness: The model should not discriminate based on protected attribute such as race, gender or age.
- Transparency: The model's decision-making process should be understandable and explainable.
- Privacy: Sensitive personal information should be protected.
- Social Impact: The model's outcomes should not exacerbate existing social inequalities.