

Guidelines for Project #3

Part I: CART - Methods and Applications

STA 522: Applied Statistical Machine Learning

Contents

Objective

Select a **new dataset** that includes:

- Both **numerical and categorical response variables**,
- At least **10 feature variables** (response variables are not counted as features).

Using this dataset, apply **Classification and Regression Trees (CART)** to address practical questions formulated based on the data.

Reporting Format

Follow the same format used in **Projects #1 and #2**.

Analytic Tasks

- **Common Analytic Tasks**
 - Formulate analytic questions derived from meaningful practical questions.
 - Assess whether the dataset contains the necessary information.
 - Perform exploratory data analysis (EDA).
 - Conduct necessary feature engineering.
 - Summarize existing methods (e.g., regression models, SVM, etc.) previously learned to address the formulated questions.
- **CART-Specific Analytic Task**
 - CART Regression:
 - * Provide a brief overview of key components before implementation.
 - * Summarize results for each step: (a) Hyperparameter tuning; (b) Final model training; (c) Predictions on test data; (d) Performance evaluation using appropriate metrics.
 - * Include visual representations where applicable.
 - CART Classification:
 - * Provide a concise explanation of major components before coding.
 - * Summarize results for each step: (a) Hyperparameter tuning; (b) Final model training; (c) Predictions on test data; (d) Performance evaluation using appropriate metrics.
 - * **Implementation guidance**, including determining the optimal cut-off probability based on an appropriate performance measure (refer to class notes for details).
 - Performance Comparison Across Models
 - * **Numerically compare** CART model performance with other models (fit the same training/testing data on alternative models).

- * Use **tables or figures** to visually compare model performance.