A TREE-BASED MODELING APPROACH TO UNDERSTANDING YOUTH DRUG USE

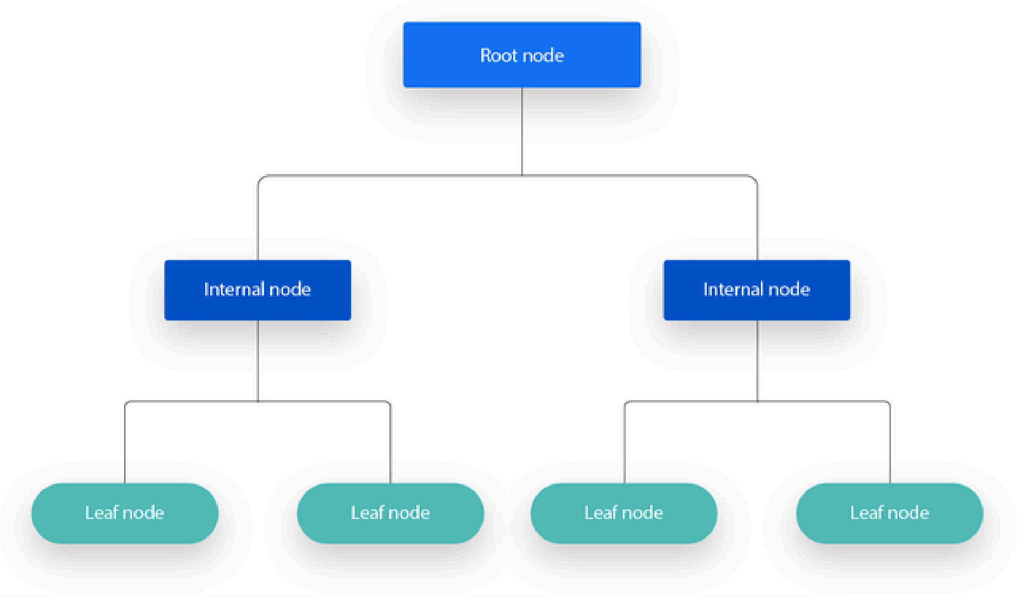
by Eloho Okoloko

Statistical Machine Learning 2 | Practical Homework 1

INTRODUCTION

- This project explores the relationship between youth substance use and demographic or behavioral factors
- The survey data used in this project was gotten from the National Survey on Drug Use and Health.
  - An annual nationwide survey administered by the Substance Abuse and Mental Health Services Administration (SAMHSA).
  - Includes responses from U.S. residents aged 12 and older, capturing information on substance use, mental health, family structure, and demographic variables.
  - past and current use of various substances, age of onset, frequency of use, household composition, and other behavioral and health-related indicators.
- Three primary questions are addressed:
  - Can we predict if a youth has ever used marijuana (binary classification)?
  - Is there a relationship between father presence and drug use history (multi-class classification)?
  - How many days has a youth consumed alcohol in the past year (regression)?
- Tree-based models used to address these questions:
  - decision trees, random forests, and gradient boosting.

THEORETIAL BACKGROUND



Decision Trees

- Intuitive, non-parametric models used for classification and regression.
- Recursively split the data based on the most informative features,
- Forms a tree where each leaf represents a predicted outcome.
- Measures:
  - Classification Error: Measures misclassification rate; simple but not ideal.
  - Gini Index: Measures node impurity; commonly used for classification - the best.
  - Entropy: Measures disorder using log probabilities; informative but computationally expensive.
- Prone to overfitting.
- Pruning—removing branches based on cross-validation performance—helps control this.

THEORETIAL BACKGROUND

Bagging (Bootstrap Aggregation)

- Builds multiple trees using bootstrapped training sets and averages their predictions.
- It reduces variance and improves stability but treats all predictors equally during splits.

Random Forests

- Extends bagging by introducing random feature selection at each split (using mtry).
- This increases model diversity and often leads to better performance.
- It is resistant to overfitting and provides variable importance metrics.
- It’s not the easiest to interpret

THEORETIAL BACKGROUND

Gradient Boosting

- Builds trees sequentially.
- Each tree fits the residuals of the previous one, learning from mistakes.
- It uses shrinkage (learning rate lambda) to control overfitting and allows tree depth tuning.
- Boosting can outperform other models but is sensitive to tuning and prone to overfitting if not regularized.
- Evaluation Metrics:
  - Classification: Accuracy and test error rate are standard metrics for model performance.
  - Regression: Mean Squared Error (MSE) quantifies prediction error.

METHODOLOGY

General Data processing and cleaning

- Dropped rows with lots of missing data
- Handled data leakage by dropping features directly related to target variable.
- Encoded categorical variables as factors
- Split dataset into training and test sets (80/20).

METHODOLOGY

Problem 1:
Binary Classification-Marijuana Use

- Buillt a classification tree
- Pruned using cross-validation to determine size(complexity) of tree
- Compared unpruned vs. pruned models using test error and gini index

Problem 2
Multi-class Classification-Father Presence

- Trained a random forest classifier using drug related variables (substance_cols
- Tuned mtry to find the best performing model
- Computed and compared model accuracies with bagging model

Problem 3:
Regression-Alcohol Use Days

- Built a gradient boosting regreesion model
- Tuning was done using different lambda values
- Compared training and test MSE across shrinkage value.

RESULT: PROBLEM 1 BINARY CLASSIFICATION

- Predict whether a youth has ever used marijuana (MRJFLAG)
- Model Used: Decision Tree (pruned & unpruned)
- Dataset: Cleaned demographic + belief variables
- Best Model: Pruned Decision Tree
- Model Test Error Rate: Unpruned Tree 13.16%, Pruned Tree 13.16%
- There might have been something wrong with my implementation

RESULT: PROBLEM 2 MULTICLASS CLASSIFICATION

- Predict father presence in the household (IFATHER with 3 classes)
- Model Used: Random Forests (mtry = 2–4) and Bagging
- Predictors: (e.g., IRMJFY, MRJFLAG, ALCFLAG, TOBFLAG, IRSEX, NEWRACE2)
- ModelAccuracy:
  - Random Forest (mtry = 2) 75.69%
  - Random Forest (mtry = 3) 75.47%
  - Random Forest (mtry = 4) 75.76%
  - Bagging Model 75.69%
- Models struggled to predict the "Don't Know" class, likely due to class imbalance.
- Strong Predictors: IRMJFY (used marijuana), ALCFLAG, and TOBFLAG

RESULT: PROBLEM 3 REGRESSION

- Predict number of days alcohol was used in the past year (IRALCFY)
- Used Gradient Boosting (with different shrinkage/λ values)
- Best Lambda (Shrinkage): 0.05
- Performance:
- Train MSE: 88,213, Test MSE: 97,973
- Even at optimal shrinkage, test error is large, suggesting the target variable is highly variable.
- I wasn't able to use the log transformed variable

DISCUSSION

Problem 1 – Marijuana Use (Binary Classification)

- Key Predictor: Importance of religion (RLGIMPT)
- Test Error: 13.16% (no gain from pruning)
- Simple models can still be effective; more behavioral features may improve accuracy.

Problem 2 – Father Presence (Multiclass Classification)

- Best Accuracy: 75.76% (Random Forest, mtry = 4)
- Top Predictors: Marijuana, alcohol, and tobacco use
- Models failed to predict “Don’t Know” class most likely due to class imbalance issue.

Problem 3 – Alcohol Use Days (Regression)

- Best Model: Gradient Boosting ($\lambda = 0.05$)
- MSE: 88,213 (train), 97,973 (test)
- High error reflects wide response range; more refined features could help.

CONCLUSION

- My models were not able to make interpretable predictions
- Both trees used RLGIMPT (importance of religion) as the key predictor.
- Youths who consider religion important were less likely to report marijuana use

REFERENCES

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: with applications in R (2nd ed.). Springer.

Lesson Notes/Slides

THANK YOU