Predicting Substance Usage

Applying tree-based methods to youth data from the

National Survey on Drug Use and Health (NSDUH)

2023 [1]

Overview: Data

- A little more than 10,000 records
- Mostly categorical data with more or less meaningful codes [2]
- Youth responses to NSDUH 2023, filtered [3]
- Gives insight into social, demographic, and behavioral factors that may be related to drug use
- As marijuana products are easily available and smokeless tobacco products make it easier in some ways for child usage to go undetected:
    - ➢ Can other data be used to identify youths who may be at higher risk to inform usage of limited outreach and health resources?

Trees

Regression

- Split based on RSS (recursive binary splitting)
- Predict based on mean
- Early splits more important
- Greedy approach can lead to high variance

Classification

- Split based on node purity
  - Gini,
  - Entropy,
  - Deviance
- Predict based on mode

Pruning

May decrease complexity and variance of the model since pruned trees are more likely to be similar.

Cost Complexity Pruning:

- Too costly to check every subtree
- Cost function penalizes depth
- Create an index of penalties and find the best tree for each penalty, then choose the penalty based on CV error

Bagging

Reduces variance through resampling and averaging

- No overfit on the number of trees but there are diminishing returns

Out-Of-Bag Error (OOB Error):

- Like CV error, estimates the test error
- Mean of errors for each tree on predicting the values not included in the bag
- OOB metric depends on response type and context

Random Forest

Start with bagging but limit the features that a tree may use:

- Decorrelation results in greater variance reduction
- Cannot overfit the number of trees
- Must tune Mtry, the number of features considered in a tree
  - Lower values will result in lower complexity and variance
  - Too low might miss important interactions
  - Tune based on OOB metric

Boosting for Tree Ensembles

Idea: learn slowly and carefully

- Trees learn from previous trees incrementally by fitting residuals

Tuning:

- Number of Trees (B): too many may overfit slowly
- Learning rate (λ):
  - Weight contributions of trees to ensemble
  - Prevents large learning steps that may increase variance
  - Too small may result inefficient learning steps and more trees needed
- Interaction Depth (D):
  - Limits the number of features and interactions allowed in a tree

Methods: Missing Data

Demographic Data:

- Codes that did not represent legitimate question skips marked NA

Substance Use Data:

- Histograms and Frequency Charts show little difference in cleaned data

After marking missing data, rows with missing data omitted

Methods: Data Transformations

Codes changed to zero to allow numeric or ordinal comparison (except age of first use features):

- 91, 93, 991, 993,
- 5, 6

Skipped School → Binary

Current or upcoming grade → Categorical

Days of Marijuana Use → log(Days of use + 0.001)

ALCYDAYS → reduced levels to none, seldom, and often

Methods: Models

Problem 1:

- Tree
- Pruned Tree
- Bagged Forest
- Random Forest

Problem 2:

- Tree
- Pruned Tree
- Random Forest

Problem 3:

- Tree
- Pruned Tree
- Boosted Forest

Methods: Tuning

Random Forest Models:

- Tuned on OOB error with randomForest::tuneRF
    - Start at the square root of the number of features and search nearby

Boosted Models:

- Tuned features based on CV score, but also kept data for best parameters based on test score

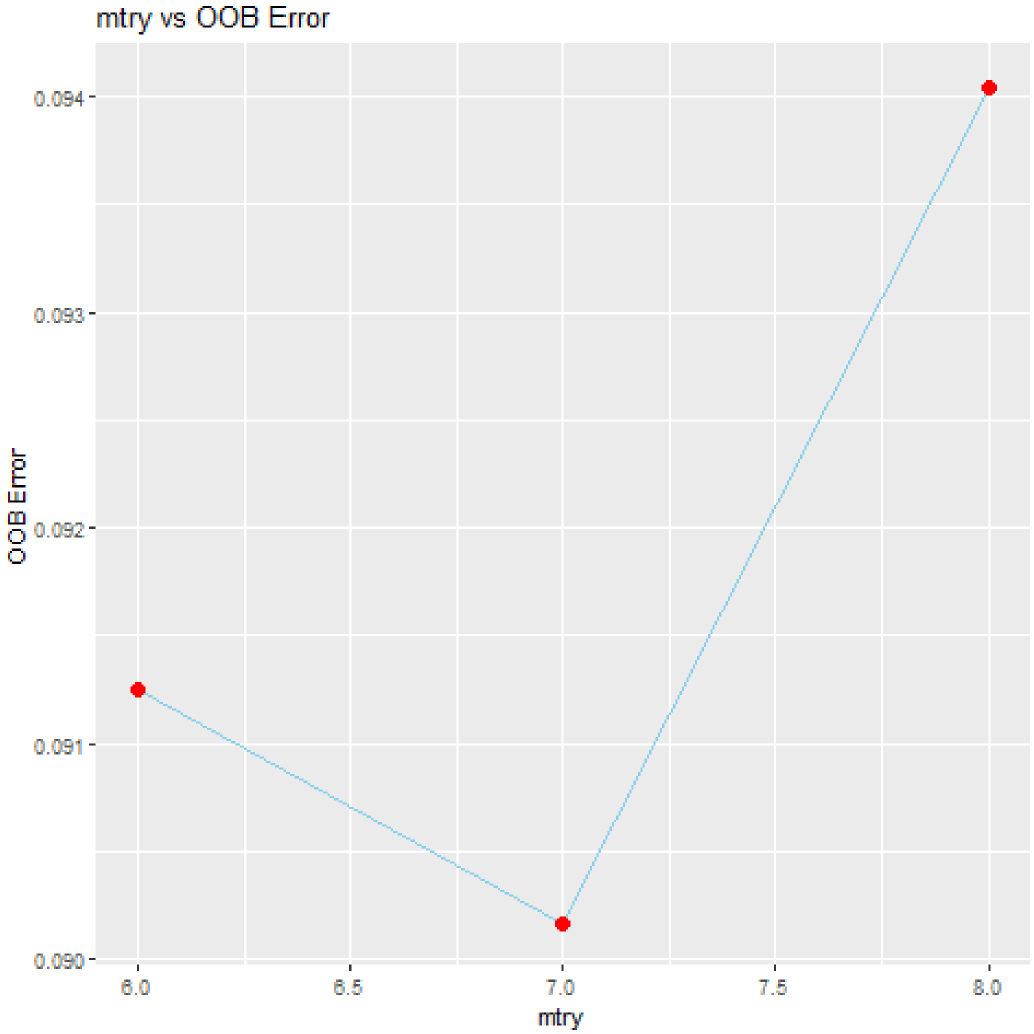
Methods: Model Evaluation Details

- Tuning was conducted using CV or OOB error on a training set
- Comparison of tuned models based on validation set
- MSE for comparison of regression models:
  - For transformed model, transformation was reversed on predictions so that training and validation error may be compared
- Classification Models:
  - Accuracy fails to measure performance on imbalanced classes
  - F1 scores for each class and weighted mean F1 do better

Problem 1: Can NSDUH data predict whether a youth has ever used tobacco products?

Results: Random Forest Tuning

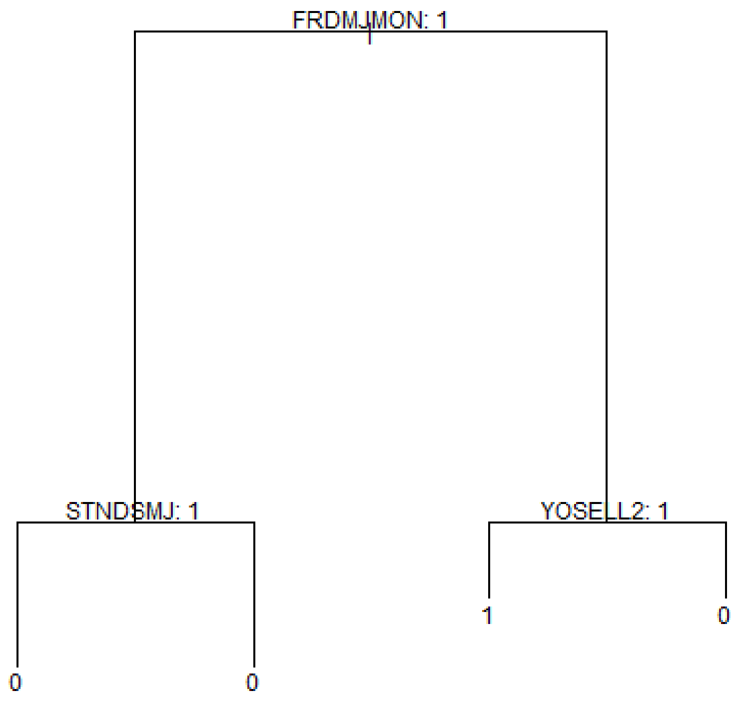
Chosen Value: 7



Results: Metrics

Model	Test Error	Balanced f1	NoTobacco f1	YesTobacco f1
tree	0.11194	0.264472	0.940239	0.117647
bagging	0.109453	0.285314	0.941255	0.2
random_forest	0.110075	0.265174	0.941294	0.119403

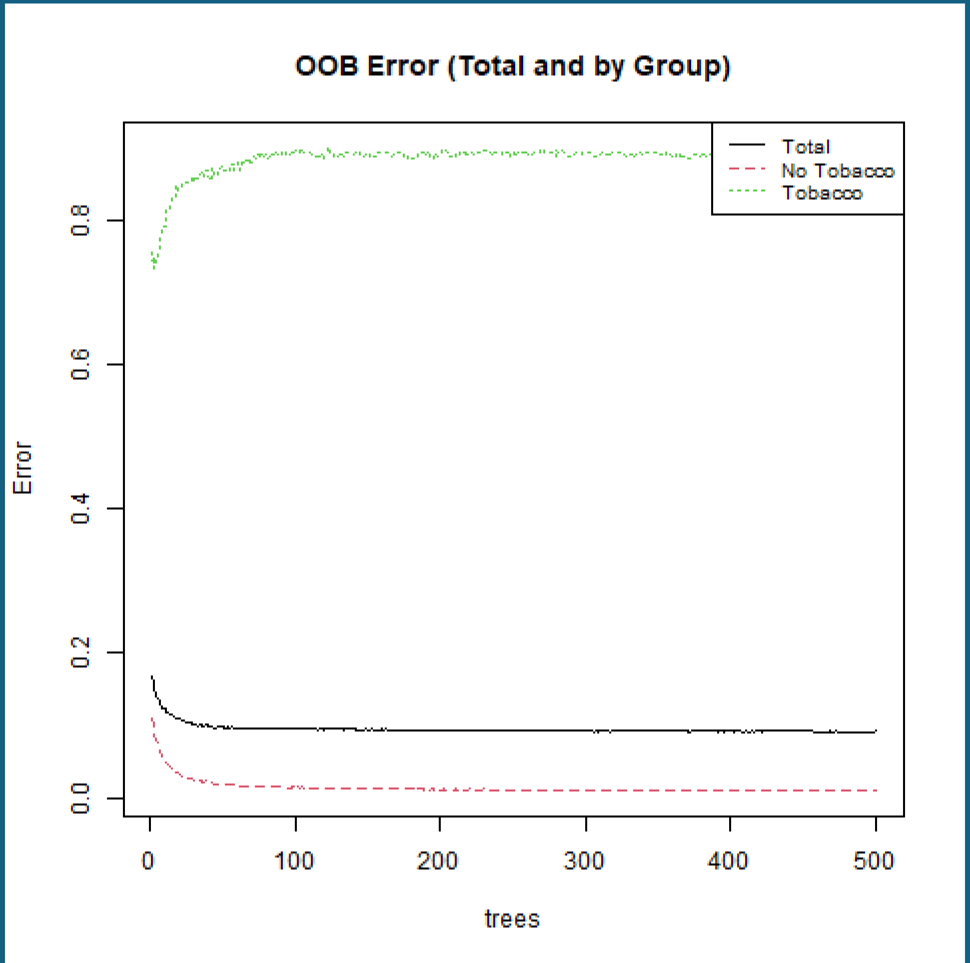
Results: Feature Importance

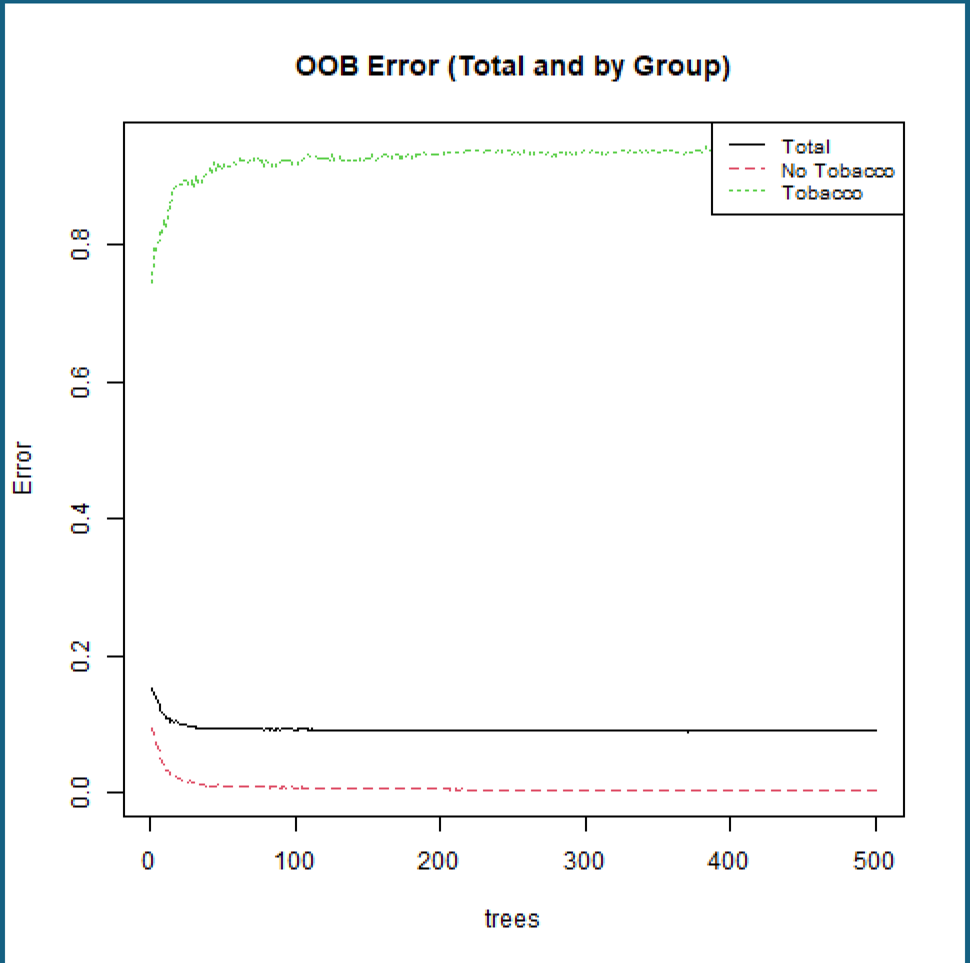


Bagging Variable Importance:

Feature	Gini Improvement
EDUSCHGRD2_T	97.4475
NEWRACE2	67.59912
HEALTH2	60.9783
INCOME	44.50357
FRDMJMON	41.43099
YFLMJMO	33.52151
YOSELL2	32.2252

Error vs N-trees for bagging (left) and random forest (right)

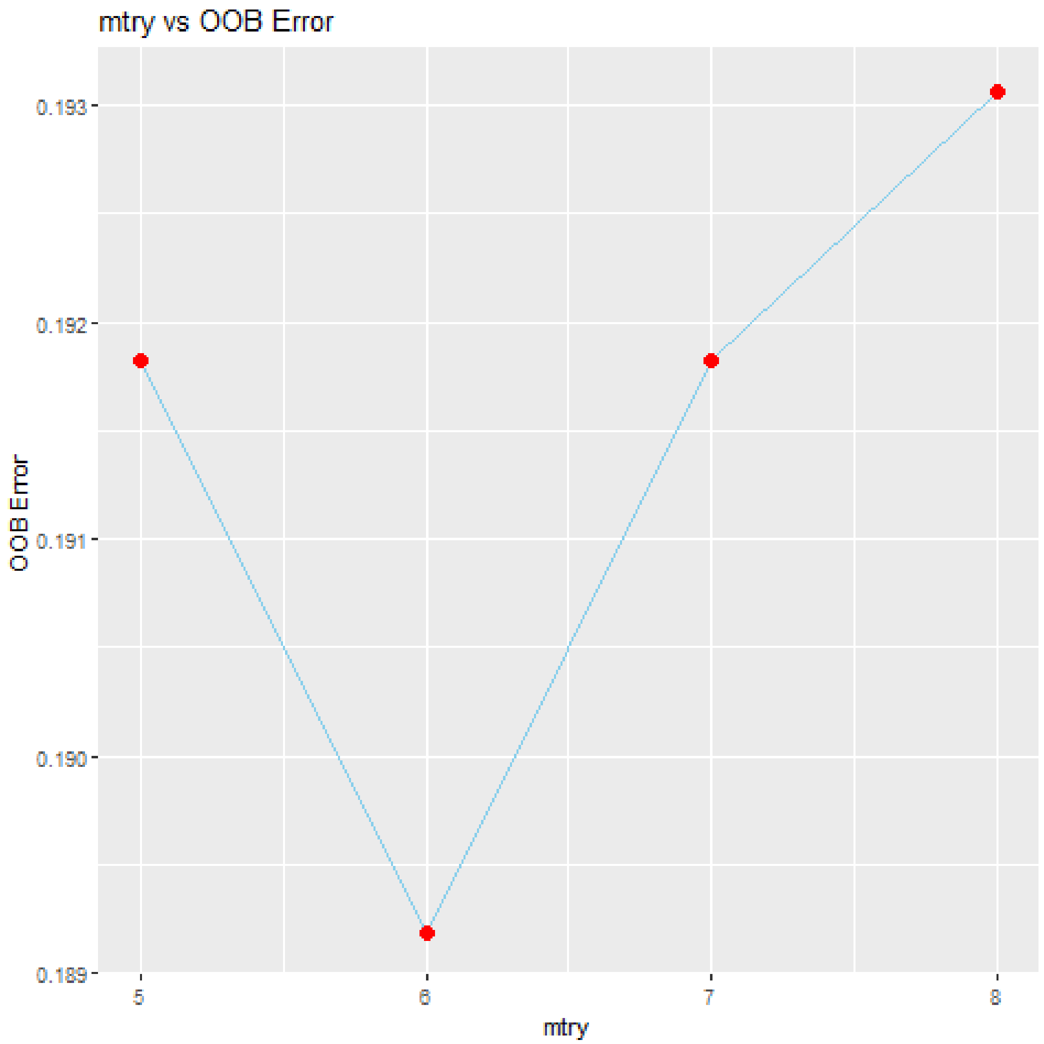




Problem 2: Can NSDUH data predict whether a youth has had alcohol never, seldom, or more often over the past year?

Results: Random Forest Tuning

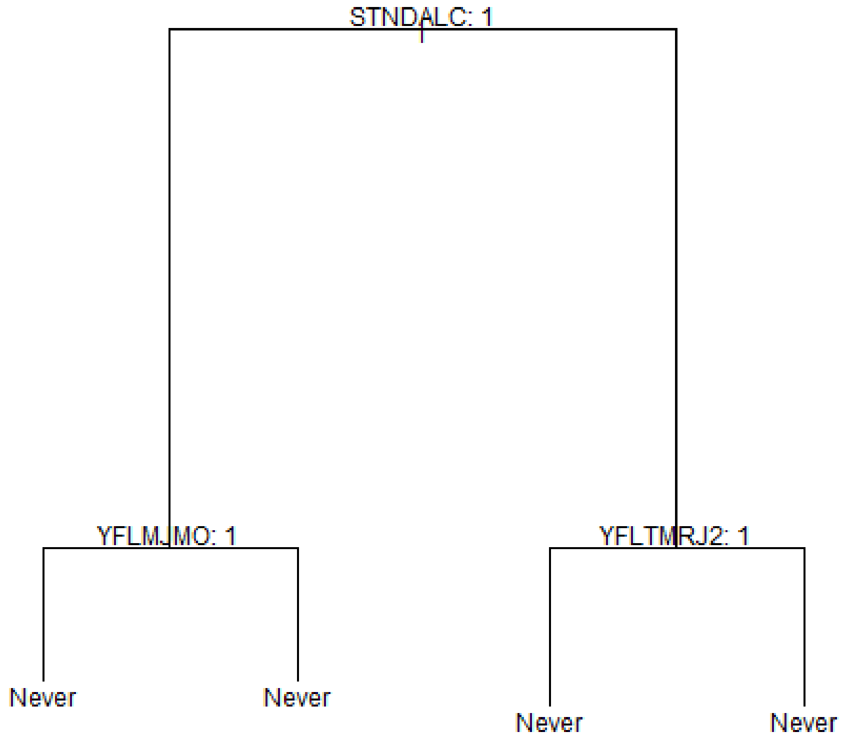
Chosen Value: 6



Results: Metrics

Model	Test Error	Balanced F1	Medium Use F1	Low Use F1	High Use F1
tree	0.196517	0.099004	0	0.891034	0
pruned_tree	0.196517	0.099004	0	0.891034	0
random_forest	0.067786	0.137292	0.2	0.900494	0.135135

Results: Feature Importance

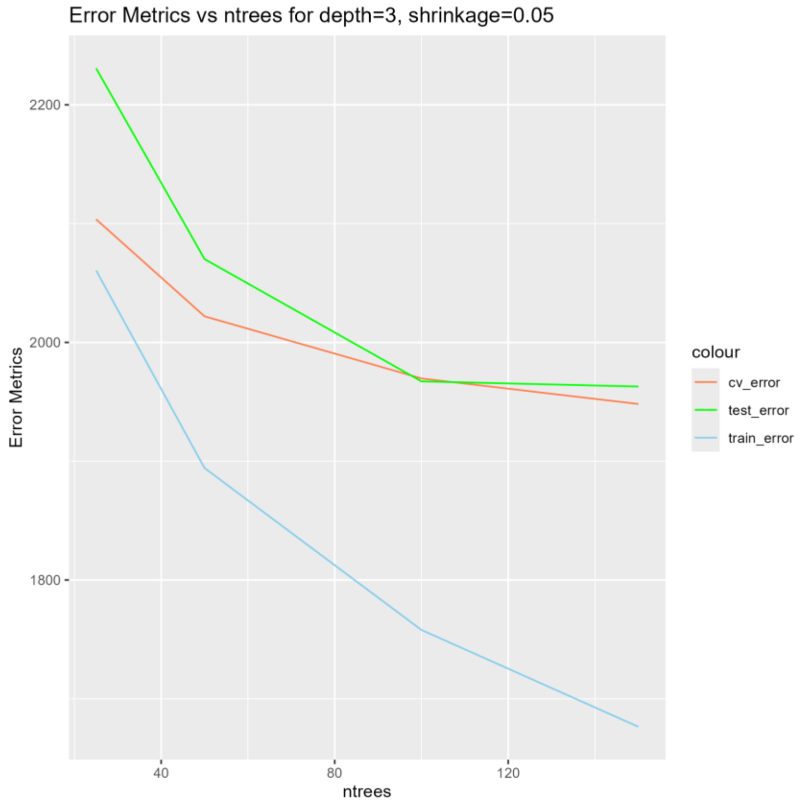


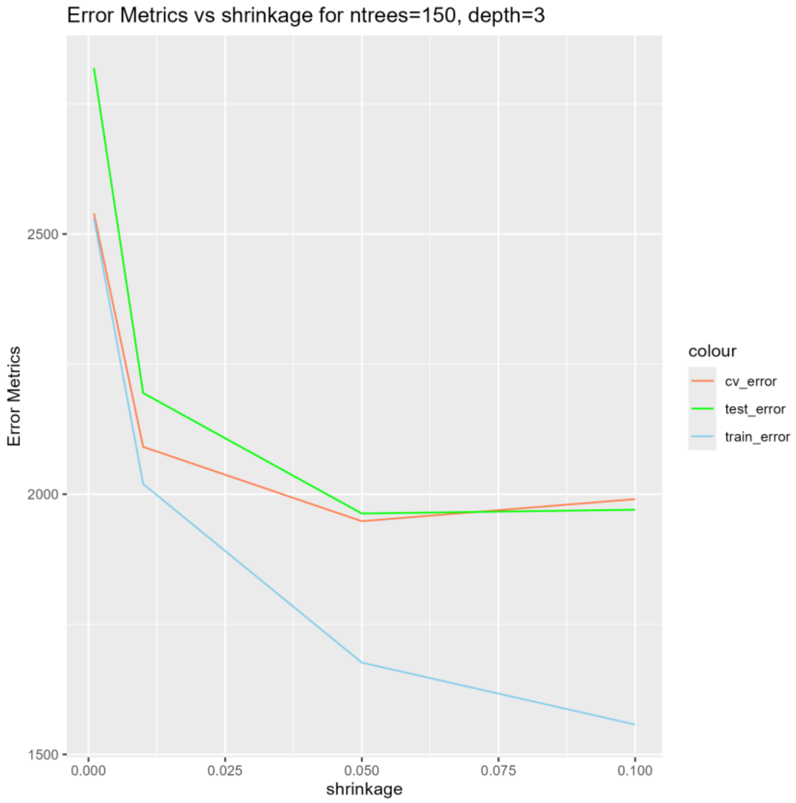
Feature	Moderate/Often	Never	Seldom	Gini Improvement
EDUSCHGRD2_T	10.87651	9.612343	8.636564	147.9477
NEWRACE2	0.429546	8.30367	2.047458	88.82681
STNDALC	24.5962	28.42077	14.85239	83.76328
HEALTH2	3.481492	2.04633	-1.57251	73.74942
EDUSKPCOM_T	1.178549	3.674957	-1.35792	58.65024
INCOME	3.447175	10.43357	2.645908	58.23583

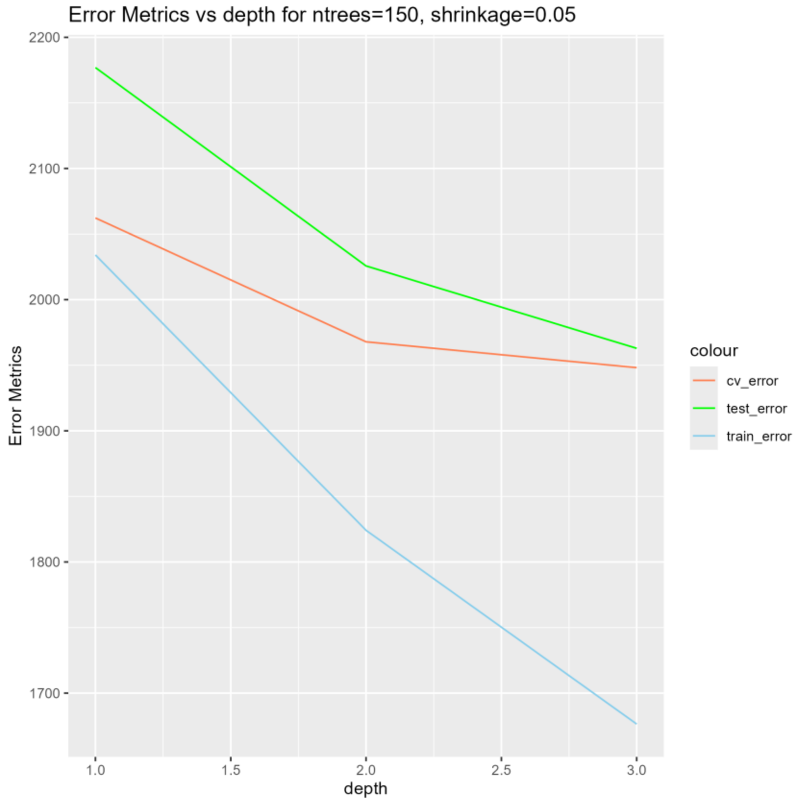
Problem 3: Can NSDUH data predict the number of days a youth used marijuana in the past year?

Results: Boosted Ensemble Tuning

Best Params based on Test MSE: (B, λ, D) = (150, 0.05, 3)







Results: Metrics

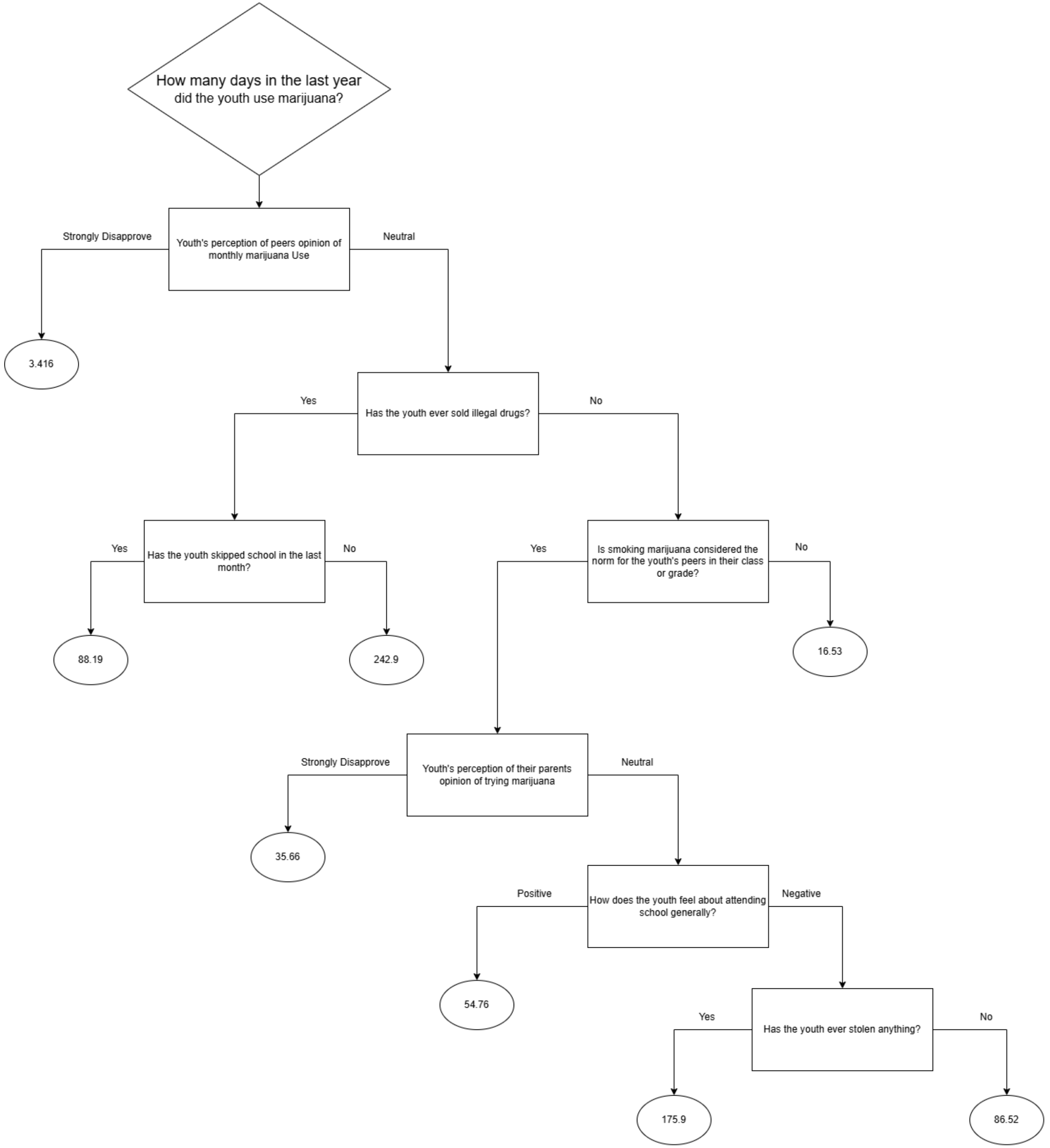
Model	Test MSE	Log Transformed Response?
tree	2101.54	FALSE
pruned_tree	2132.539	FALSE
tree	2942.799	TRUE
pruned_tree	3232.882	TRUE
boosting	1976.785	FALSE
boosting	3199.137	TRUE

Results: Feature Importance

Feature	Relative Influence
YOSELL2	15.18155
FRDMJMON	13.81026
STNDSMJ	8.434671
EDUSCHGRD2_T	6.978973
PRMJMO	6.18905
YFLMJMO	6.172565
YOSTOLE2	5.441146

A youth:

=> Whose peers do no feel strongly about monthly marijuana use
=> Who has not sold drugs
=> Whose peers often use marijuana
=> Whose parents are not strongly against marijuana use
=> Who feels poorly about school
=> And who has stolen before
=> Predicted use is very high, 176/365



Further Discussion: Choosing the Response

- Predicting on the categorical versions depends on class imbalance
  - Can give a general idea of usage
  - May be more appropriate when classes result in more balance than the numeric version
  - Reducing number of classes may reduce class imbalance
- Prediction on a numeric response depends in part on normality
  - Still can only predict a discrete set of values, unlike a linear model

Further Discussion: Variable Importance

Race as a predictor:

- Important to models in part 1 and 2
- Correlation is not causation
- Use of race may lead to discrimination or bias feedback loop

Criminal history as a predictor:

- Likely a causation link → may be useful
- Could lead to a feedback loop

References

[1] Center for Behavioral Health Statistics and Quality. (2024). *2023 National Survey on Drug Use and Health (NSDUH),* Substance Abuse and Mental Health Services Administration. Rockville, MD

[2] Center for Behavioral Health Statistics and Quality. (2024). *2023 National Survey on Drug Use and Health Public Use File Codebook,* Substance Abuse and Mental Health Services Administration. Rockville, MD

[3] Mendible, Ariana. (2025). *5322* [source code]. GitHub. https://github.com/mendible/5322

[4] R Core Team (2025). *R: A language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

[5] Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, **4**(43), 1686. doi:10.21105/joss.01686.

References

[6] Ripley, B.D. (2023). *Tree: Classification and Regression Trees*. R package version 1.0-43. https://CRAN.R-project.org/package=tree

[7] Liaw A, Wiener M (2002). "Classification and Regression by randomForest." *R News*, **2**(3), 18-22. https://CRAN.R-project.org/doc/Rnews/.

[8] Ridgeway, G., Greenwell, B., Boehmke, B., GBM Developers. (2024). *gbm: Generalized Boosted Regression Models*. R package version 2.2.2. https://CRAN.R-project.org/package=gbm

[9] James, G. , Witten, D., Hastie, T., Tibshirani, R. (2023). *An Introduction to Statistical Learning with Applications in R*. Springer https://www.statlearning.com