Predicting Substance Usage

Studying 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?

Overview: Methods

- **Simple trees** for comparison and interpretability
- **Bagging** improves the result by reducing variance (resampling)
  - **Random Forest** takes it one step further
    - Randomly limiting the features considered limits outsized effects during training
    - Penalizing complexity through number consider => more general
- **Boosted Tree Ensemble:**
  - Also penalizes complexity through learning rate, interaction depth, and ensemble size
  - Adds idea that trees should focus on the relationships missed by previous trees, training on the residuals and weighting by contribution
  - Powerful, but prone to over-fitting in ensemble size
- **Tools:** R [4], tidyverse [5], tree [6], randomForest [7], gbm [8]

Missing Data

- Codes that did not represent legitimate question skips marked NA
- Initial Chi-Squared and Kolmogorov-Smirnov tests suggest possible differences in IRALCFY, IRALCAGE, and ALCYDAYS after cleaning, but…
  - Histograms and Frequency Charts show little discernible difference
  - Kolmogorov-Smirnov can be unreliable for non-interval variables

Data Transformations

- Substance use codes that mean zero use => zero
  - 91, 93, 991, 993,
  - 5, 6
- Demographic codes that are not useful => NA

- Treating days of school skipped and grade as categorical

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

Results

Tree:

- Error Rate: 0.112
- Pruned Tree assigned all to never-smoker status

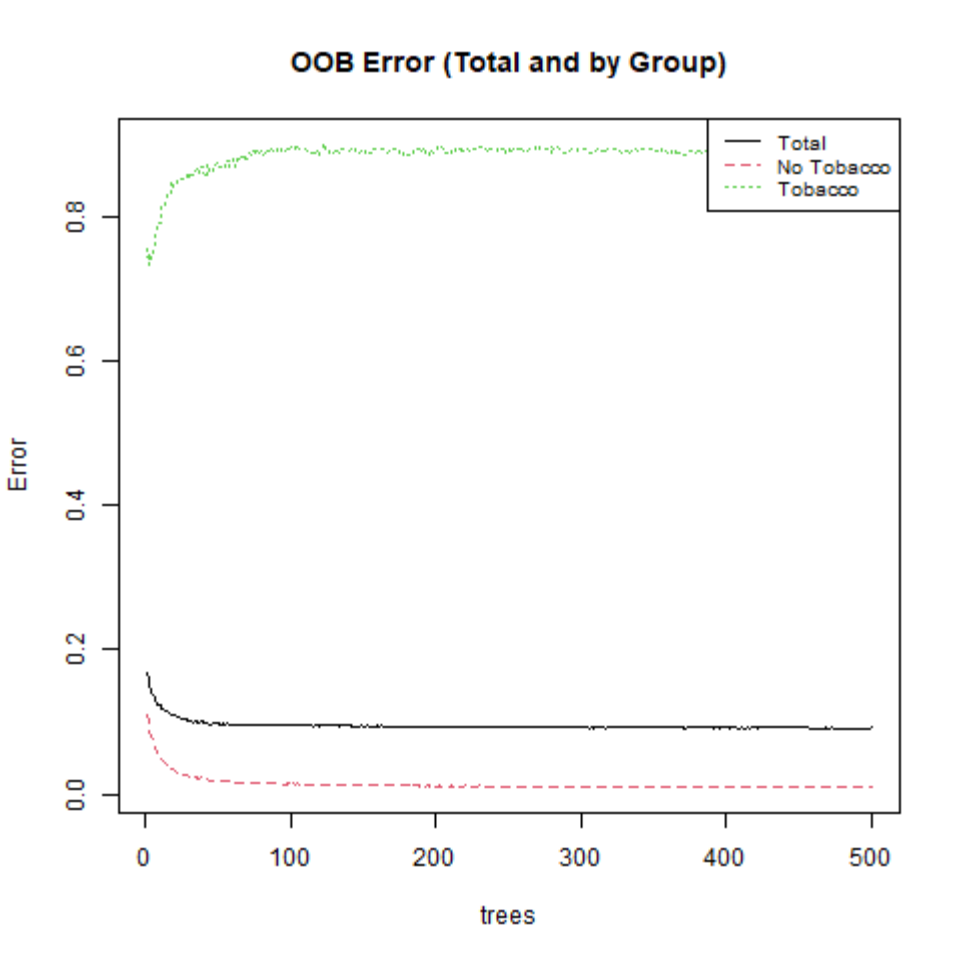
Bagging:

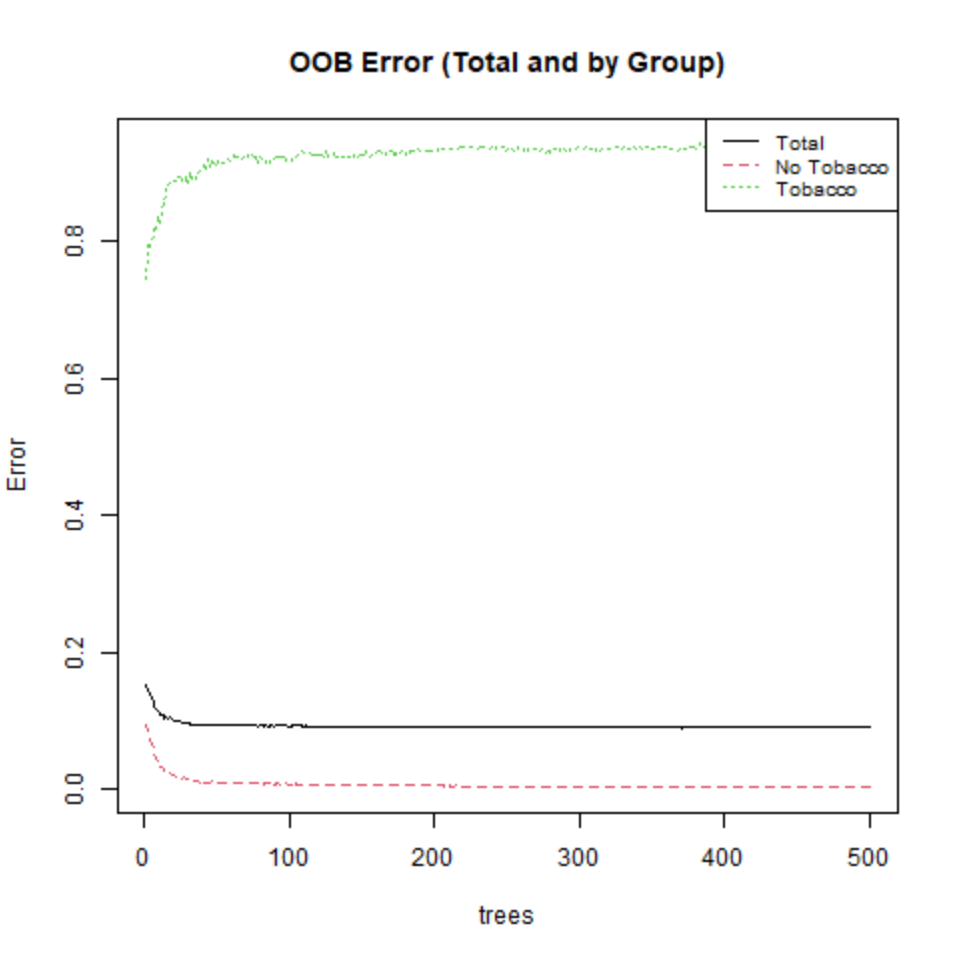
- Error Rate: 0.109
- Slightly more nuance but error rate for smokers still high

Random Forest:

- Error Rate: 0.11
- M-features retained chosen as 7, but varies
- Still fails to reliably identify smokers

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





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

Results

Tree:

- Error Rate: 0.081
- Both pruned and full tree assign all never-smoker status

Random Forest:

- Error Rate: 0.068
- M-features retained chosen as 6, but varies
- Similarly, as ensemble size increases, smoker error gets close to 1

* Use of gbm in R not appropriate as no multinomial version exists

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

Model Results

Tree:

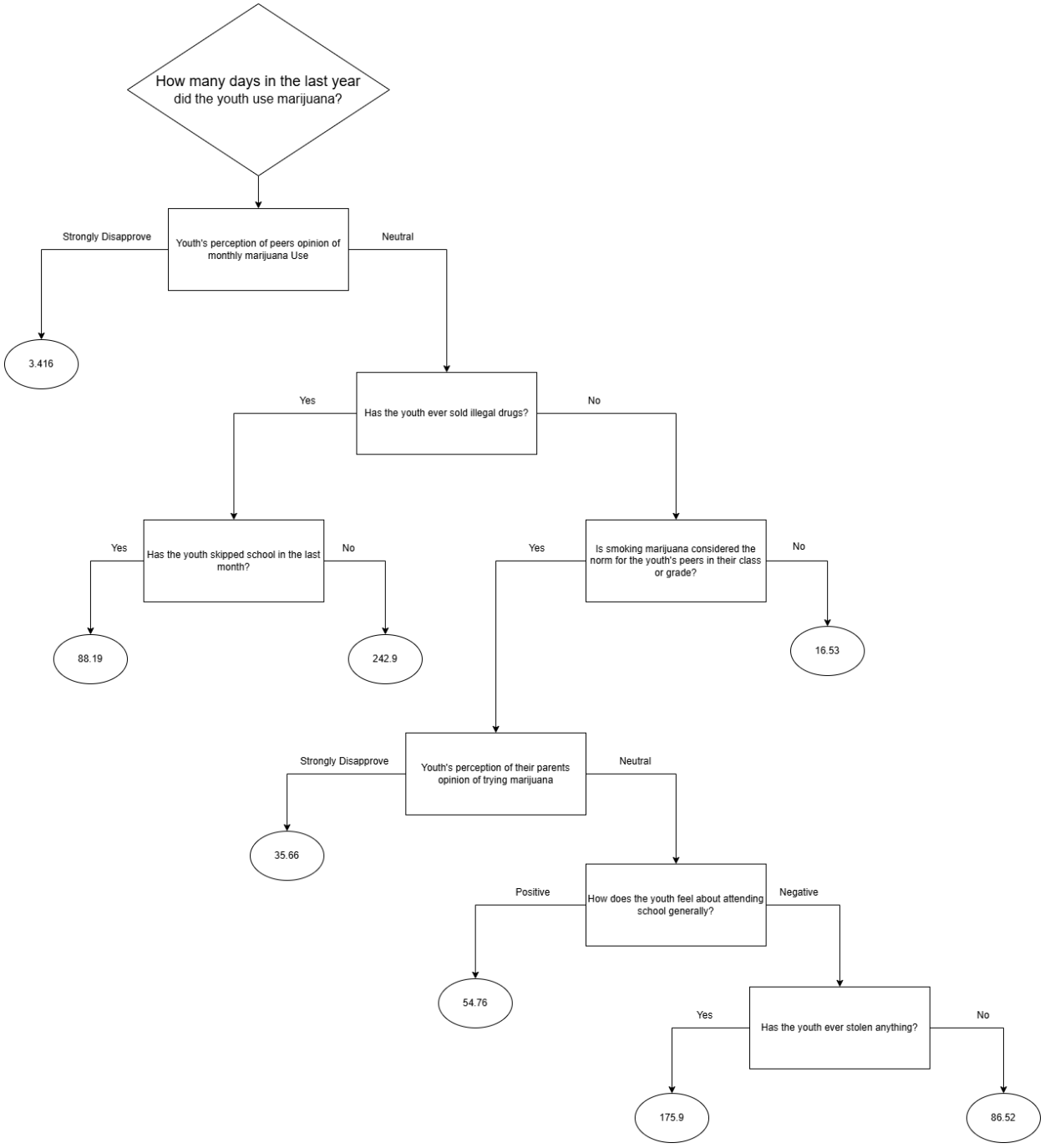
- MSE: 2101.54
- Much more interesting full and pruned decision tree

Boosted Tree Ensemble:

- MSE: 1991.95, CV.Error: 1944.25, Train MSE: 1759.4
- Chosen Parameters:
  - Ensemble Size: 100
  - Depth: 3
  - Shrinkage: 0.05

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



What Features are Important?

With the exception of the boosting model in problem 3, a number of features maintain relative importance between the other models: grade, race, health level, income level, perception of peers and ones own opinion on marijuana, the sale of drugs, skipping school, poverty status, county population level, the ubiquity of marijuana use, history of theft, and census area population level

For the boosting model: Some of the same features were important, but in a different order.

- History of selling drugs, opinion of peers, and ubiquity of use , most important

Conclusion

The models had some success, with ensemble methods obtaining fairly low test error each of the problems. However, the bagging and random forest methods tended to have very high error for users in problem one and seldom users in problem 2. To a degree, error was also high for the heavy users in problem 2. These are just the sort of youth we want to detect, limiting the usefulness of the models. Still some interesting factors were identified. Social pressures appear to have a major impact or relationship worth looking into. It's also worth exploring difference in urban and rural areas, and in economic groups.

Limitations and Future Exploration

- Models may benefit from over and undersampling skewed data
- More data about the feelings of students towards various social issues might be interesting: perhaps data from social media could be used to gain further insights to tune the models
- More rigorous testing is required to ensure proper treatment of missing data codes, since ALCYDAYS was used as a target

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