Social Effects on Drug Use

Jesse Loi

Introduction

- The National Survey on Drug Use and Health (NSDUH) contains data relating general features like education to drugs use.
- We will investigate how social factors may impact marijuana use in youth
    - Social factors: opinions of peers, parental support

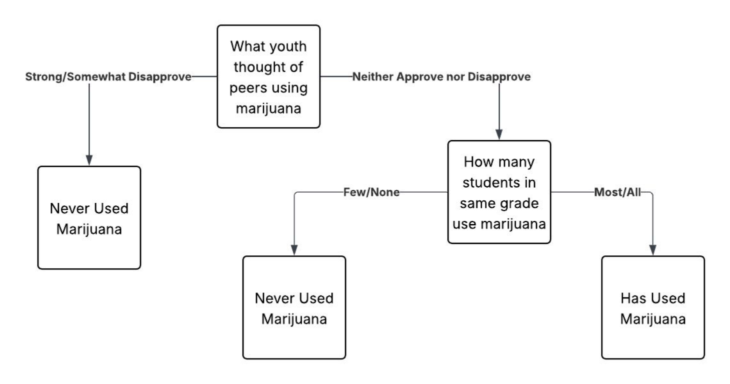
Models:

- Simple classification and regression decision trees
- Bootstrapped Aggregated (Bagging) Decision Trees
- Random Forest Decision Trees
- Boosted Decision Trees

Goal:

Minimize Mean-Squared Error and Misclassification Rate

Theoretical Background: Trees



Simple Decision Tree

Bootstrap Aggregating:

Re-sample data many times and build many trees

Flaw: one impactful variable may dominante (try Random Forest)

Random Forest:

Avoids dominance of a single feature by growing many trees with different sets of features

Boosting:

Train our data on errors of previous data

Flaw: tends to overfit (cross validate)

Theoretical Background: Models

Relevant Features: # of trees

Bootstrap Aggregating:

Re-sample data many times and build many trees

Flaw: one impactful variable may dominante (try Random Forest)

Random Forest:

Avoids dominance of a single feature by growing many trees with different sets of features

Relevant Parameter: # of Features

Boosting:

Train our data on errors of previous data

Flaw: tends to overfit (cross validate)

Theoretical Background: Evaluation Metrics

$$\text{Precision} = \frac{\text{correctly classified actual positives}}{\text{everything classified as positive}} = \frac{TP}{TP + FP}$$

$$\text{Recall (or TPR)} = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{TP}{TP + FN}$$

$$\text{F1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{2TP}}{\text{2TP} + \text{FP} + \text{FN}}$$

Accuracy Scores:

Mean-Squared Error (MSE): For regression tasks (estimating a numerical value)

Misclassification Error: For classification tasks (which percentage of cases we classified correctly)

Other Scores:

Precision: Proportion of positives out of all labeled positives (including false positives)

-Use case: security clearance

Recall: Takes the correctly classified positives out of all positives (including false negatives)

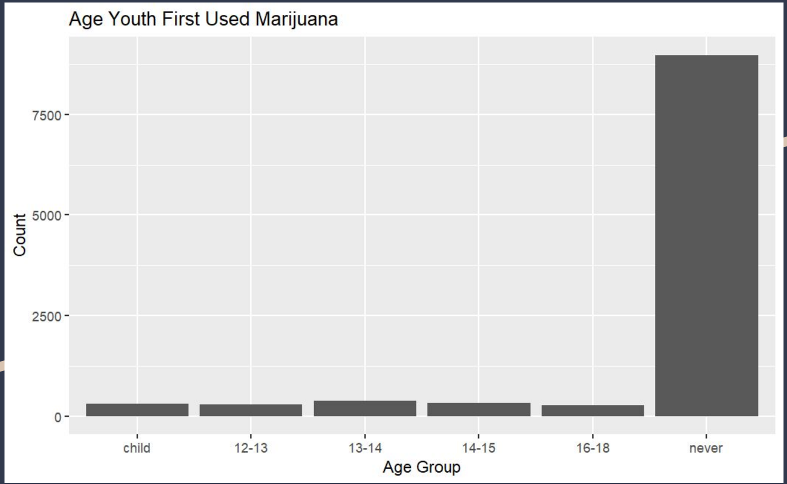
-Use case: disease detection

F1 Score: an average of these two scores

Methodology: Cleaning

Len : 3 ALCOHOL FREQUENCY PAST YEAR - IMPUTATION REVISED

RANGE = 1 - 365 ....
991 = NEVER USED ALCOHOL ....
993 = DID NOT USE ALCOHOL PAST YEAR ....



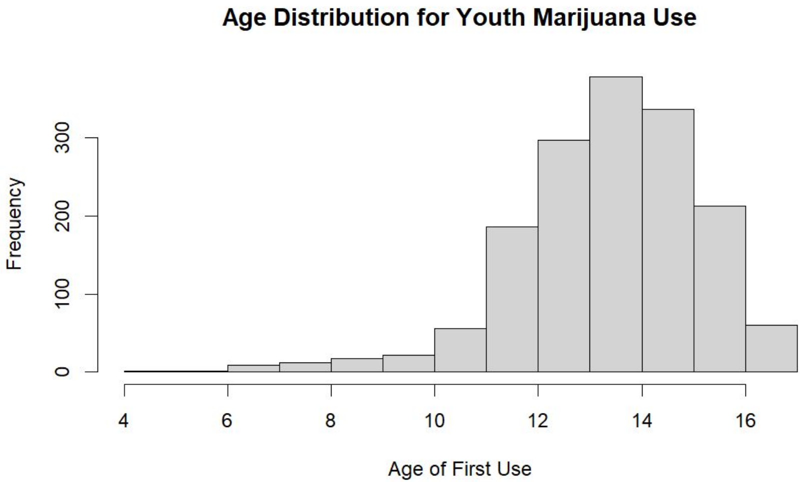
<u>Data Cleaning</u>

- Convert variables to categorical
- Change Extreme Values to 0

<u>Problem: Unbalanced Data</u>

Could lead to models w/ high accuracy but low precision

Solution: omit “never” entries and investigate only those who have used marijuana



Methodology: Feature Selection

Other Features:

Days in previous year smoking tobacco

Days drinking alcohol past year

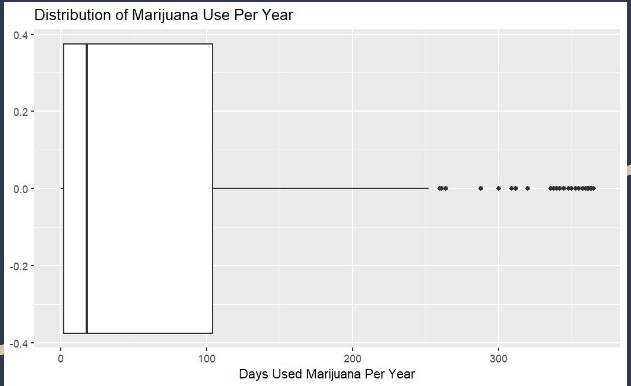
Age first used tobacco

Age first used tobacco

“Social” Features

- Teacher told the student they did a good job
- Gradēmates use marijuana
- Parents tell youth they did a good job
- Parents tell youth they are proud
- Opinion of close friends smoking more than 1 pack a day
- Youth talked to parent about alcohol, tobacco, and drugs
- Youth participated in a self esteem group
- Participated in a substance prevention program
- Participated in a program to help with substance use
- Youth sees drug prevention messaging outside school
- Youth had drug education in school
- What the youth thinks of peers using marijuana monthly)

Methodology: Questions for Each Model Type



Binary Classification:

How can we classify youth into having used marijuana and never having taken marijuana.

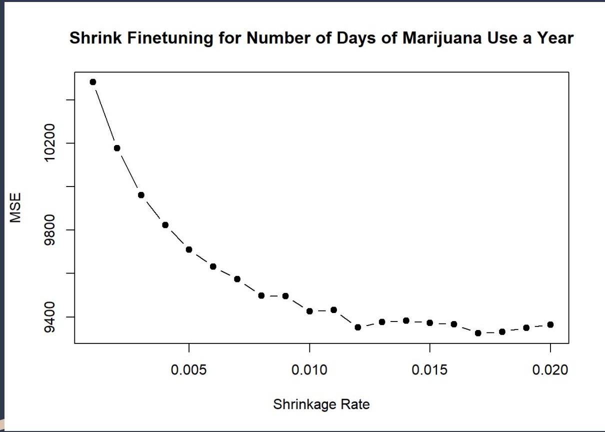
Multiclass Classification:

How can we determine when a marijuana user first used marijuana?

Regression:

Can we predict how many days a year a youth will have used marijuana?

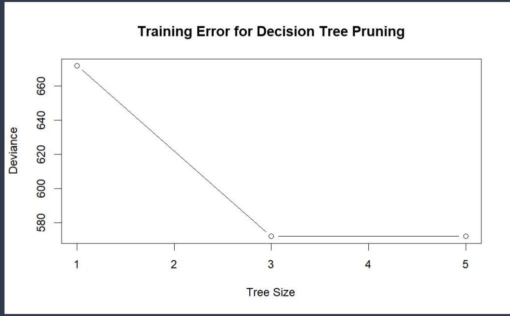
Methodology: Hyperparameter Tuning

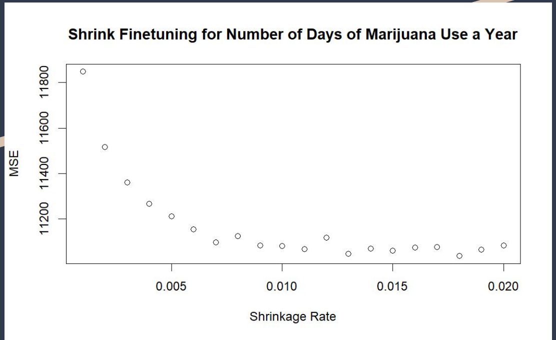


Finetuning Procedure

1. Run a loop through all the levels I want to check
2. Create a different model for the range of hyperparameters I want to check
3. Take the relevant metric (cross validated MSE)
4. Plot the error
5. Look for an elbow point (if monotonically decreasing metric) or a minimum (if not)
    a. Look for an elbow point to balance gain with resources

Methodology: Finetuning (Simple, Boost)





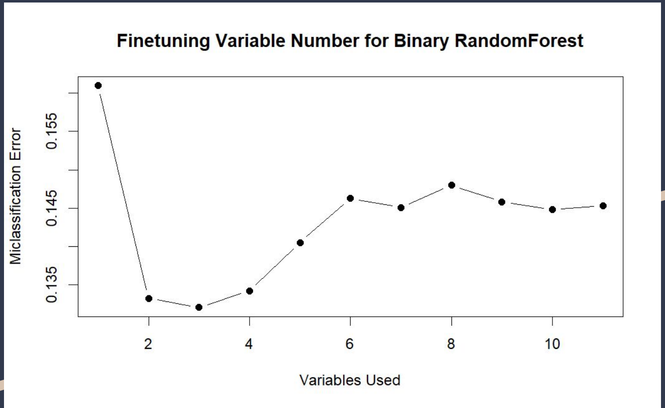
<u>For simple trees</u>: try out different tree sizes to avoid overfitting

- A tree size of 3 was the best
- No need to decrease further

<u>For Boosting</u>: we change the shrinkage rate, which is the rate at which our model updates to accommodate errors.

- Increasing the shrinking parameter did not decrease test MSE by much, but we chose an elbow point of 0.011

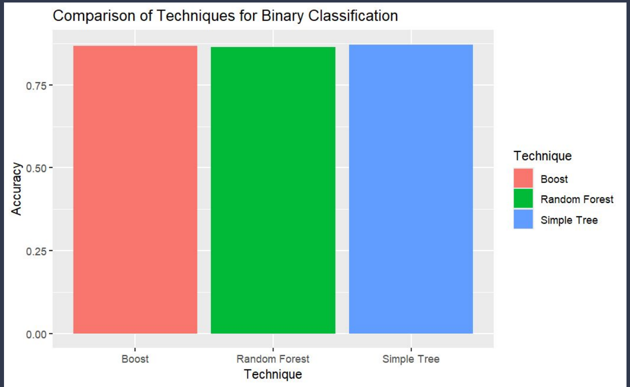
Methodology: Finetuning (Bagging, Random Forest)

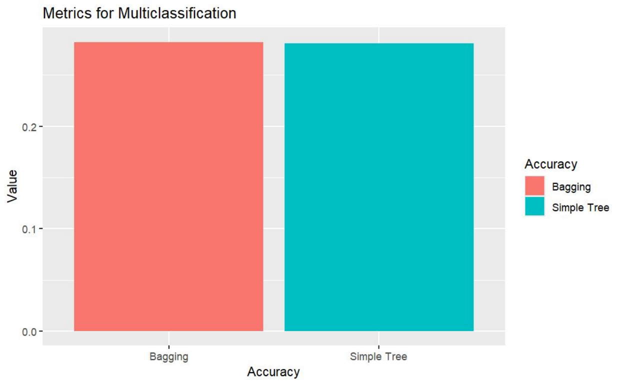


For Random Forest:

We finetune the number of variables used, ending with 3, which is very close to the size of the tree for our simple tree used before.

Results





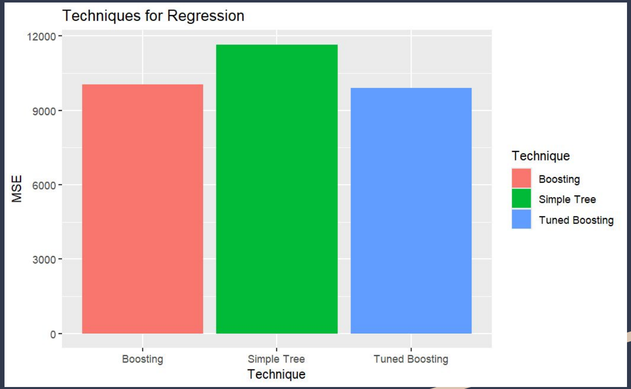
<u>Binary Classification</u>:

- Predicts marijuana use with an error rate of 12.4%
- Important Variables: opinion of peers' marijuana use, number of peers using marijuana

<u>Multiclass Classification</u>:

- Predicts age of first marijuana use with an error rate of 70%
- Important Variables: age of alcohol first use

Results

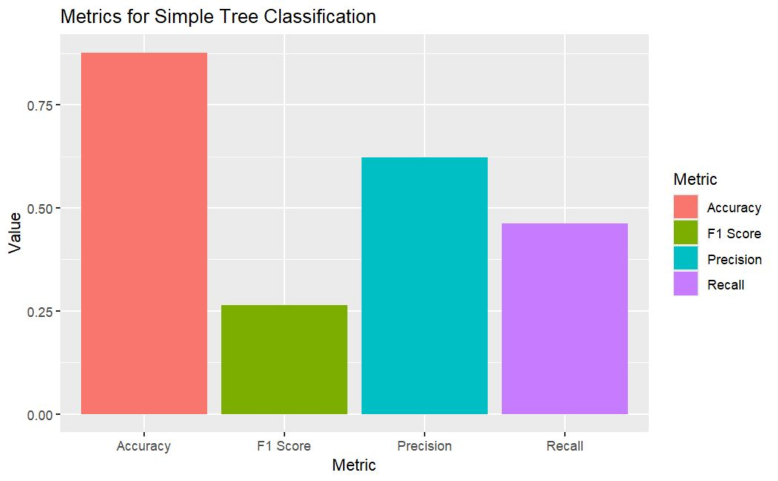


<u>Regression</u>:

- Predicts with a mean squared error of 10000 days.
- Important Variables: days per year using alcohol, whether parents say they are proud, and monthly cigarette usage

Results: Consequences of Unbalanced Data

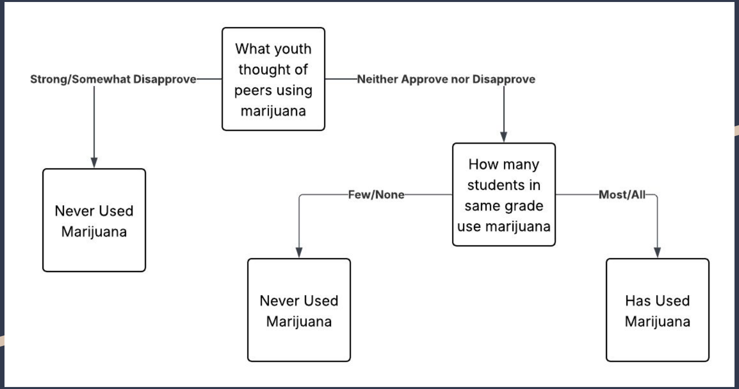
Metrics for Simple Tree Classification



- Precision high: we are not predicting many false positives
- Recall low: we are failing to classify some marijuana users

Discussion: Social Effects Based on Variable Type

A simple decision tree



Binary Classification:

- Tree size of 3 is enough (3 terminal nodes)
- Social factors related to peers are the most impactful

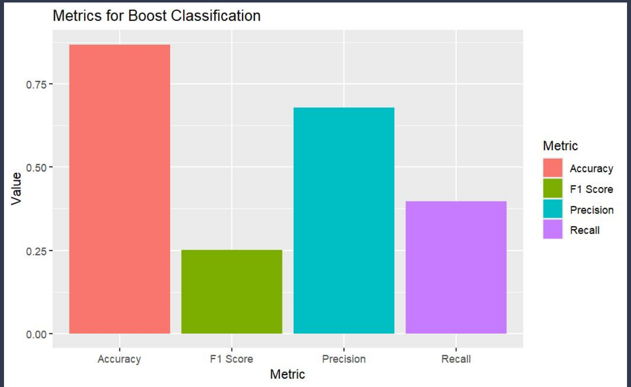
Multiclass Classification:

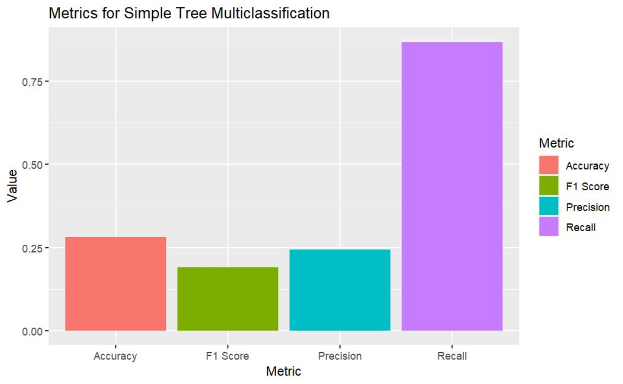
- Very difficult to determine when youth started smoking by age group.
- Most impactful factors relied on collinearity with other drug use variables (generally unreliable).

Regression:

- Frequency of drug use per year all can predict each other
- Parental intervention surprisingly helpful

Discussion: Metrics





Unbalanced data leads to lower recall.

Why is recall (for ages 14-15) so high?

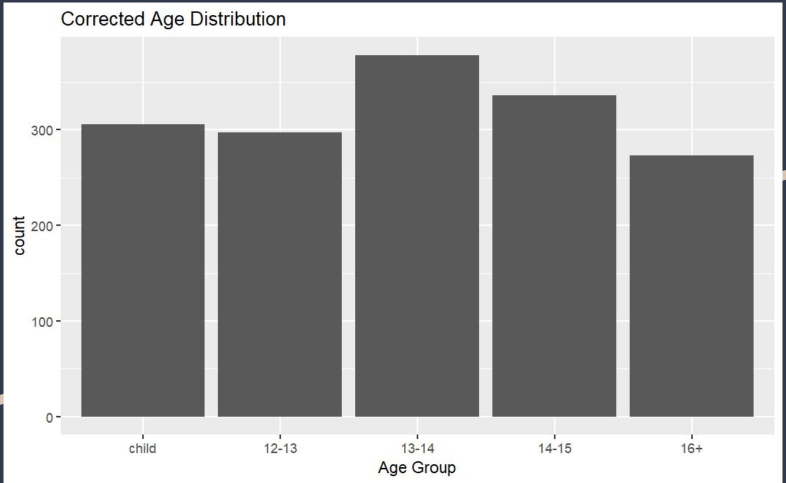
- Slightly more users at that age range led our basic tree to predict it fairly evenly.
- Varies based on the multiclassification target

Which metric is most important?

- If we want to identify ALL marijuana users (aggressively) then recall is important
- If we want to avoid labeling and be conservative, precision

Discussion: Binary, Ordinal, and Numerical Variables

Ideal for treating as ordinal



Example: Days used marijuana during the past year

Binary: categorize days into those above 182 and under or equal to 182, entailing “high” or “low” usage

Ordinal: categorize days into buckets

Numerical: have variables from 0 to 365 representing

When to use each:

Binary: when you have little or unbalanced data

Ordinal: when there aren’t enough data for regression

Numerical: when there is plenty of data to work with

Conclusion

Peer programs may be far more effective at preventing marijuana use in youth

Hard to identify risk factors for when a teen will start to use marijuana (could be many ages)

- More work can be done to collect data to predict the age.

High usage of one drug tends to imply high usage of another drug.

- Can make anti-drug campaigns more transferable
- Parents cannot stop children from first using drugs, but can lower the frequency

References

1. Ripley B (2024). _tree: Classification and Regression Trees_. R package version 1.0-44,<https://CRAN.R-project.org/package=tr>.
2. Liaw A, Wiener M (2002). “Classification and Regression by randomForest.” _R News_, *2*(3),18-22.<https://CRAN.R-project.org/doc/Rnews/>.
3. Ridgeway G, Developers G (2024). _gbm: Generalized Boosted Regression Models_. R package version 2.2.2, <https://CRAN.R-project.org/package=gbm>.
4. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
5. Google. "Accuracy, Precision, Recall & F1 Score." Google Developers, https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall