CHRISTOPHER YANG

DATA5322

Breaking Bud: Data-Driven Insights Into Youth Marijuana Use

Introduction

Question: What factor(s) are most associated with youth marijuana use? (peer, family, demographic, health, education, etc.)

Goal: Investigate which factor(s) are most predictive of marijuana use in youth under 18 years of age.

Data

- National Survey on Drug Use and Health (NSDUH) 2023 Youth Subset [1]
- Respondents aged 12-17 years old
- 79 variables
- ~10,000 respondents

Approach

Binary classification: Has the respondent ever used marijuana?	· Decision tree · Bagging
Multiclass classification: How many days has the respondent used marijuana in the past 30 days?	· Random forest · Decision tree
Regression: How many days has the respondent used marijuana in the past year?	· Gradient boost · XGBoost

Theoretical Background

Model	How It Works	Pros	Cons
Decision Tree	Splits data into regions by feature threshold that reduces impurity (Gini or Entropy)	Easy to train and interpret	Can easily be overfit without pruning
Bagging Classifier	Trains multiples trees on bootstrapped samples and combines results by majority vote or average	Improves model stability. Cannot overfit	Dominant features might overshadow others
Random Forest	Same as bagging, but adds random subset of features at each split	Improves generalization. Cannot overfit	Less interpretable due to many trees
Gradient Boosting	Train shallow trees in sequence where each corrects previous one's errors	Reduces bias through sequential learning	Sensitive to noise and overfitting if learning rate is too high
XGBoost	Extension of gradient boosting with regularization, parallelized trees, and missing value handling	Accurate and handles overfitting well	Complex tuning and difficult to interpret

Methodology

Data cleaning

- Imputed missing values with mean and mode
- Renamed variables to be more readable
- Mapped ordinal labels to midpoints for regression

Variables

- Picked 16 variables of interest across peer, education, family, etc.

Hyperparameters

- Tuned using GridSearchCV to find optimal hyperparameters

Metrics

- Classification:
  - Accuracy, precision, recall, f1
- Regression:
  - R squared, RMSE

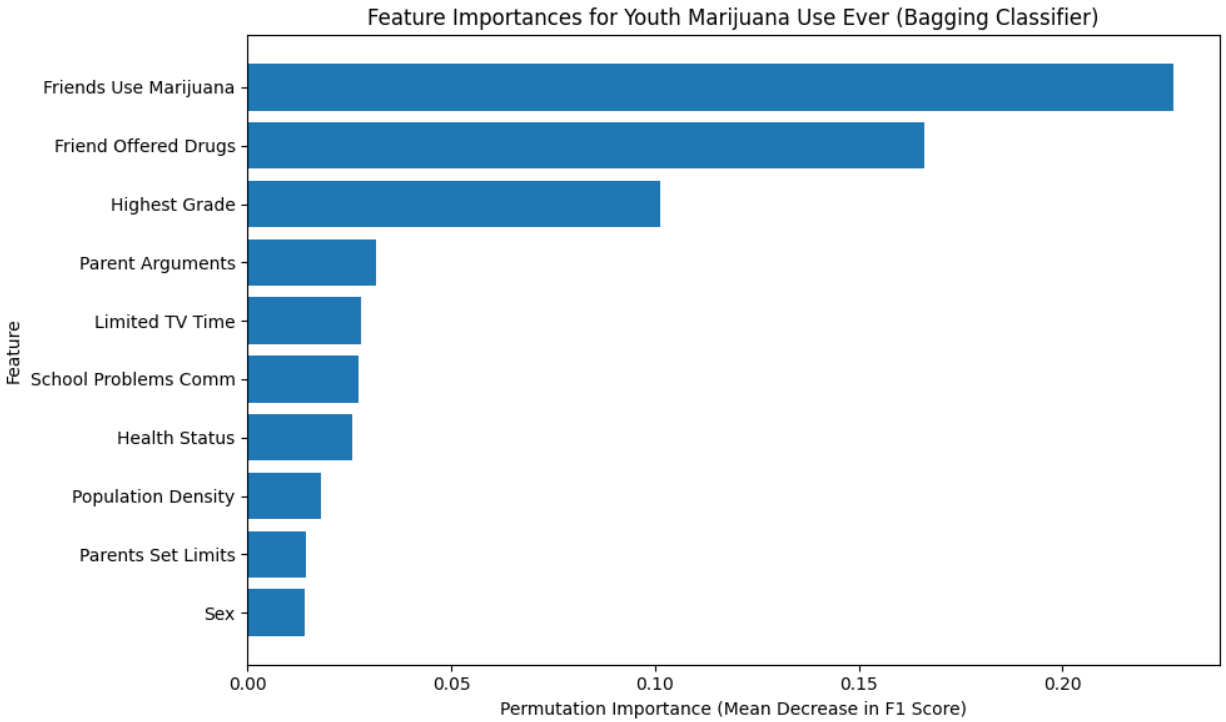
Feature Importance

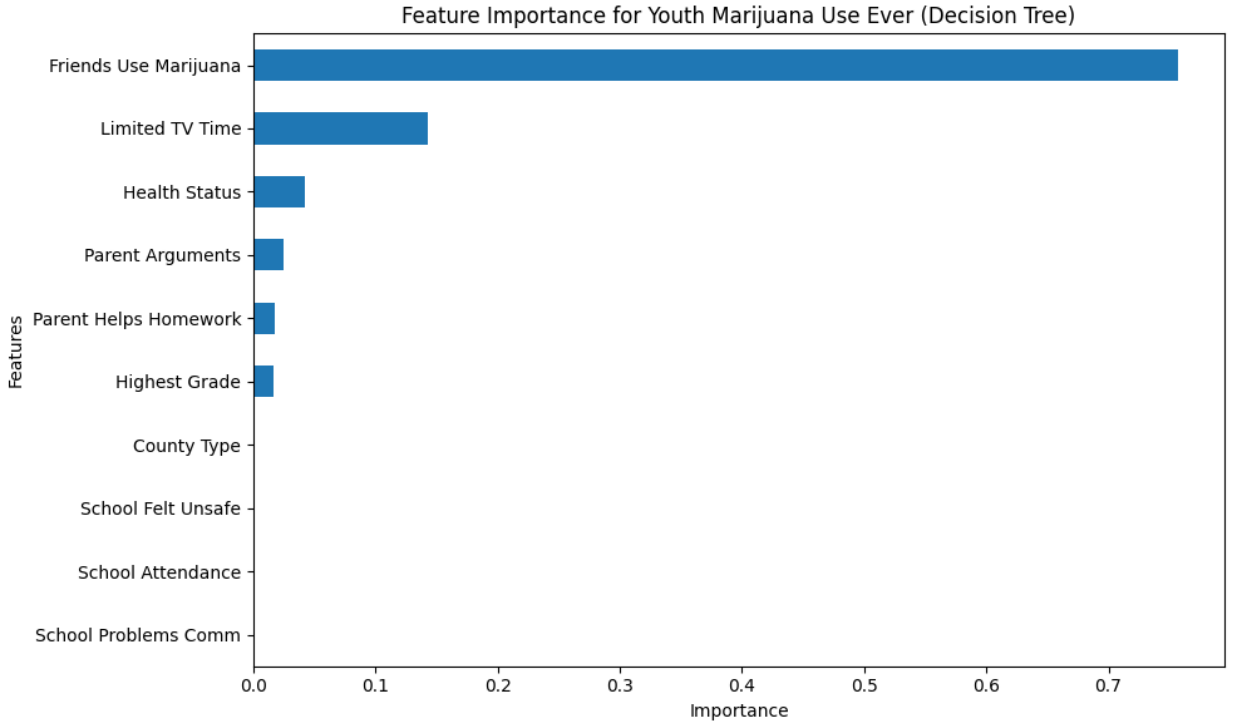
- Compare feature importances across models to find commonality

Binary Classification Results

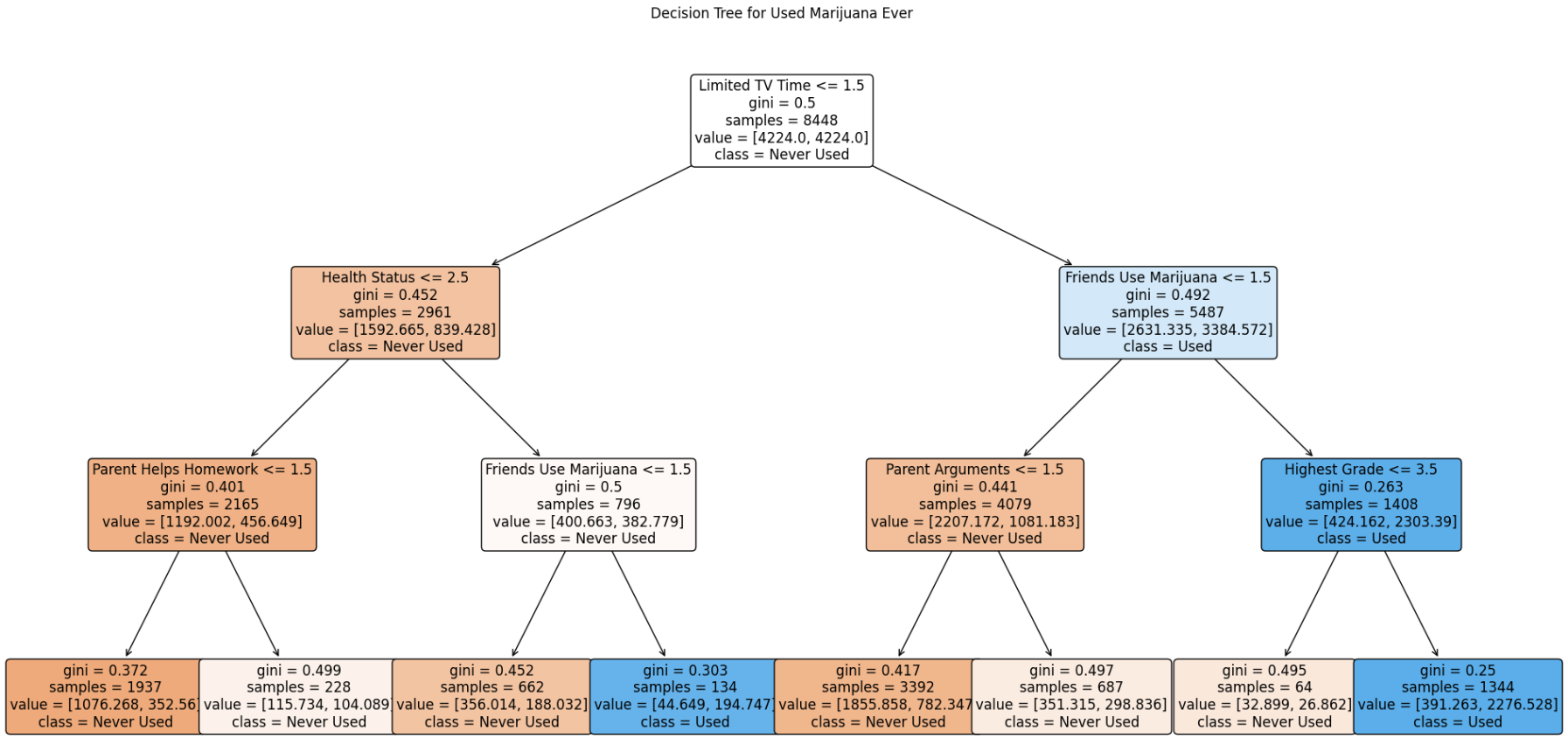
Model	Class	Accuracy	Precision	Recall	F1-Score	Samples
Decision Tree	Never Used	0.85	0.92	0.90	0.91	1781
	Used		0.52	0.60	0.55	332
Bagging	Never Used	0.86	0.88	0.97	0.92	1781
	Used		0.62	0.30	0.41	332

Binary Classification Feature Importance





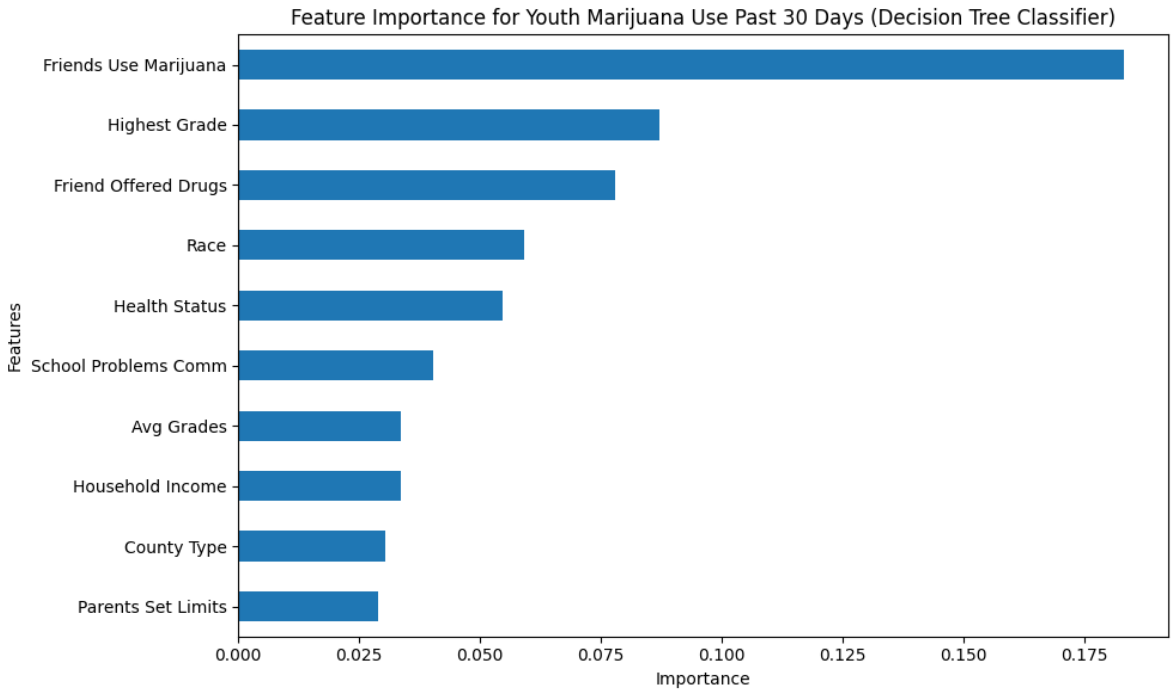
Best Binary Decision Tree

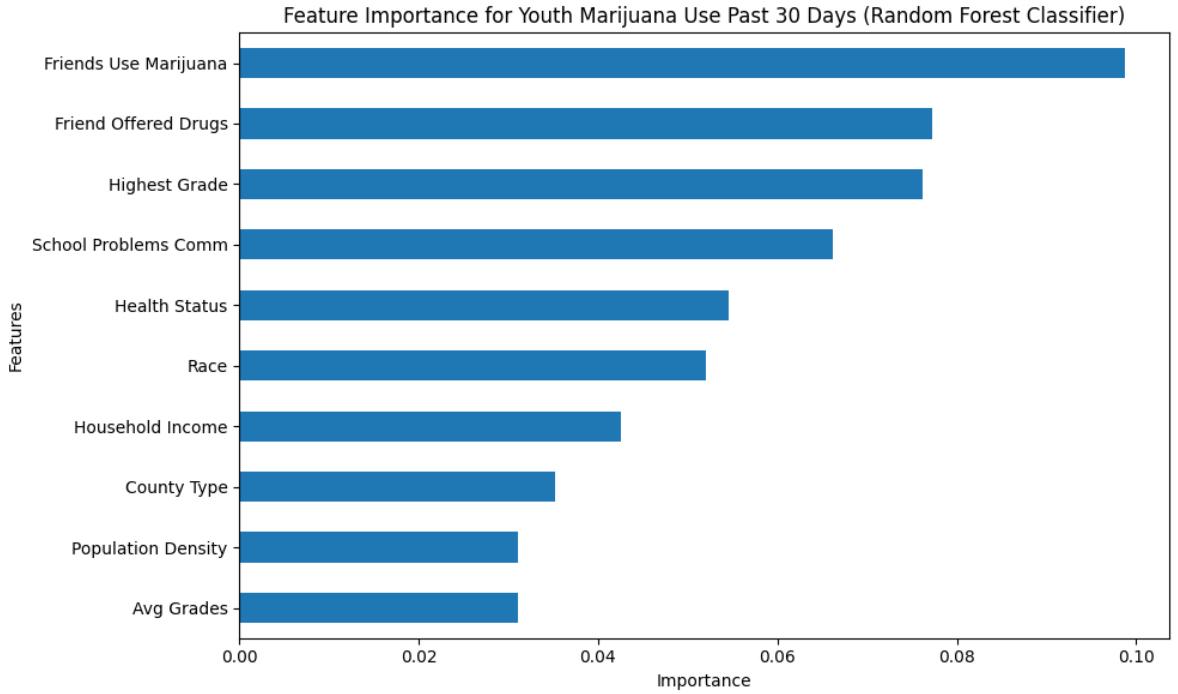


Multi-class Classification Results

Model	Class	Accuracy	Precision	Recall	F1-Score	Samples
Random Forest	Never	0.92	0.93	1.00	0.96	1954
	Sometimes		0.00	0.00	0.00	81
	Addict		0.50	0.03	0.05	78
Decision Tree	Never	0.91	0.93	0.98	0.96	1954
	Sometimes		0.00	0.00	0.00	81
	Addict		0.21	0.12	0.15	78

Multi-class Classification Feature Importance

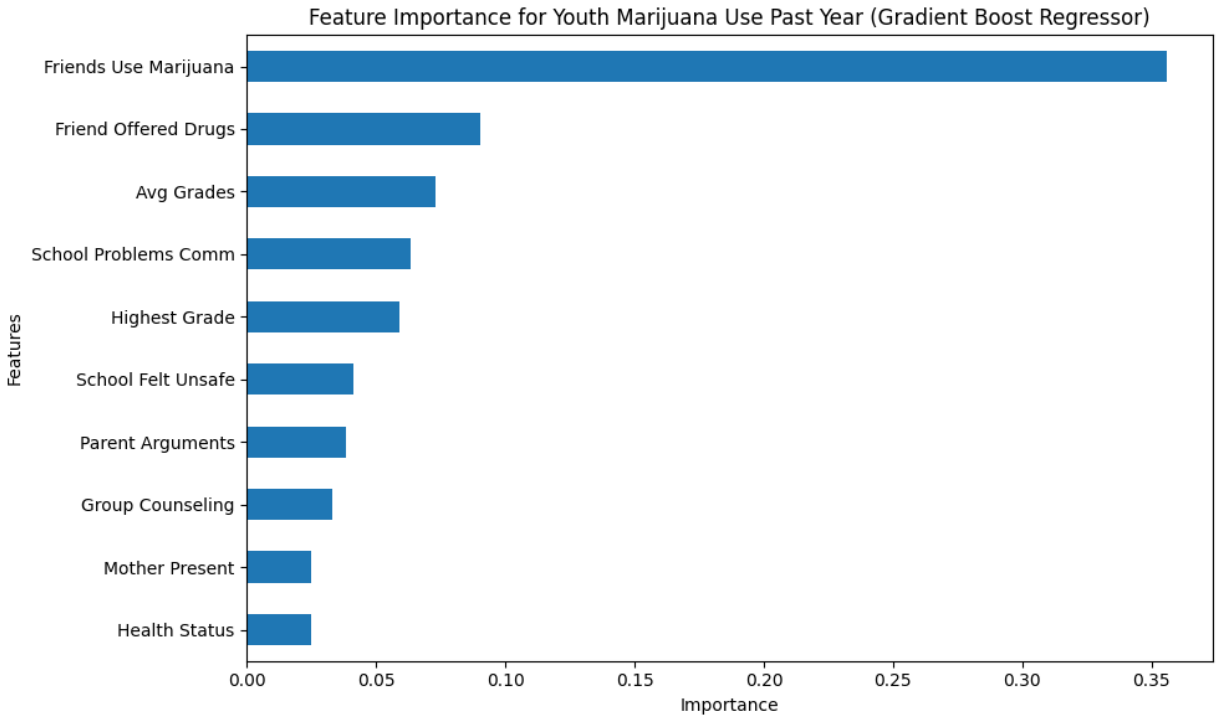


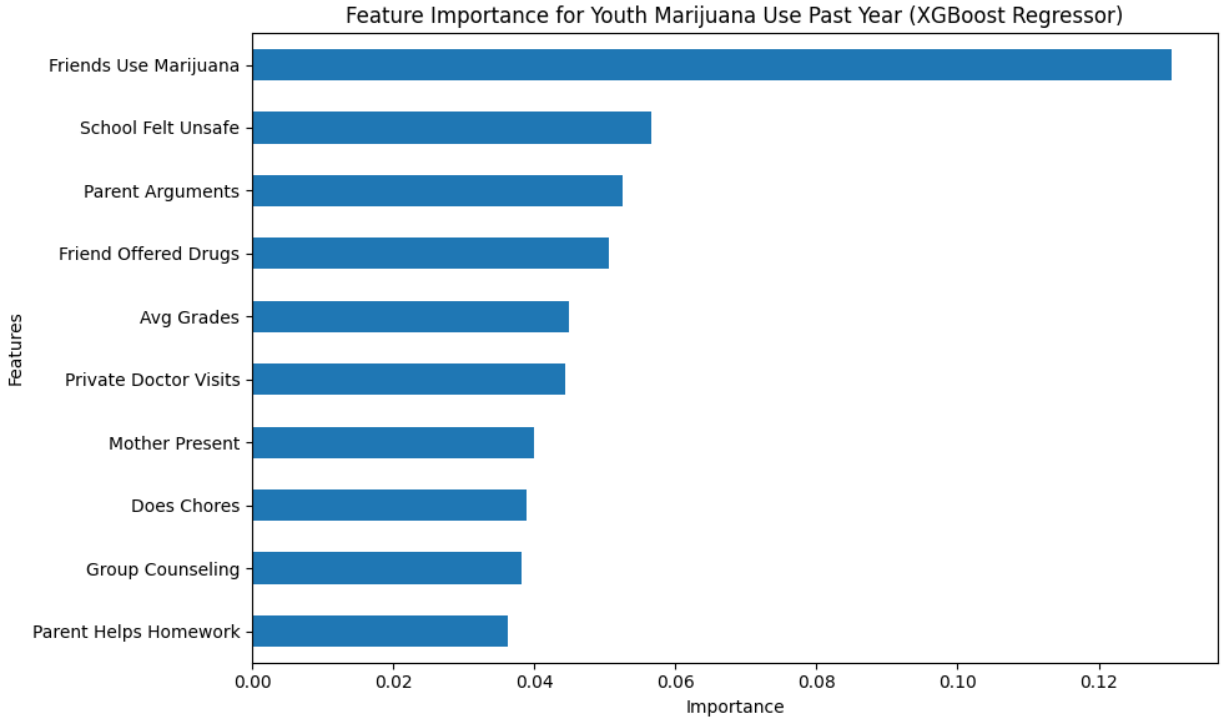


Regression Results

Model	R^2	RMSE
Gradient Boost	0.1356	41.906
XGBoost	0.1361	41.894

Regression Feature Importance





Discussion

Model	Type	Accuracy	F1 (Used/Addict)	Macro F1	CV F1 mean	CV F1 STD	Runtime
Decision Tree	Binary	85%	0.55 (Used)	0.73	0.53	0.03	~24s
Bagging	Binary	86%	0.41 (Used)	0.66	0.39	0.02	~30s
Random Forest	Multi	92%	0.07 (Addict)	0.35	0.90	0.004	~5m 18s
Decision Tree	Multi	91%	0.17 (Addict)	0.41	0.90	0.004	~10s

Discussion

Model	R^2	RMSE	CV R^2 Mean	CV R^2 STD	CV RMSE Mean	CV RMSE STD	Runtime
Gradient Boost	0.136	41.91	0.162	0.018	45.41	2.47	~6m 45s
XGBoost	0.135	41.91	0.167	0.013	45.28	2.66	~1m

Conclusion

1

Peer pressure/influence was the most influential, but not the only factor associated with youth marijuana use.

2

Class imbalance

3

How to communicate findings

References

1. Center for Behavioral Health Statistics and Quality. *2020 National Survey on Drug Use and Health Public Use File Codebook*. Substance Abuse and Mental Health Services Administration, 28 Oct. 2021, https://www.samhsa.gov/data/system/files/media-puf-file/NSDUH-2020-DS0001-info-codebook.pdf.