# Predicting Drug Abuse Susceptibility

Rishab Ghose

# Introduction

**Let's look at the facts:** Among Americans....

- **700K Drug Overdose Deaths**
  Since 2000
- **95K Deaths from Alcohol**
  Every Year
- **20M individuals ages 12 and older**
  Are affected by substance abuse disorders
- **358K adolescents**
  Had SUD and MDE within last year

# Early Awareness, Education & Prevention is KEY!

# Goal: Build a model to Predict Drug Abuse

**Substances of Interest:**

Alcohol, Cocaine, Benzodiazepines

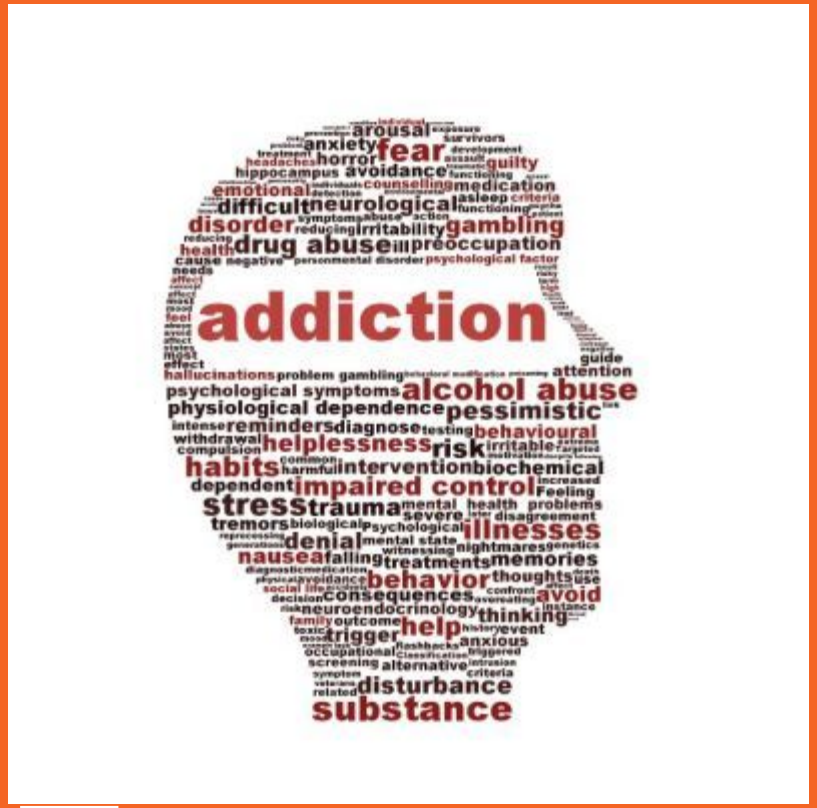
Will be using Drug Use Dataset from UCI Machine Learning Repository

# The Data

Collected from Anonymous Survey Data

Reputable Psychology Questionnaires

1885 respondents
32 Features



# Data Wrangling

- **Semeron: Fictitious Drug**
  Used to rule out untruthful respondents
- **7 Personality Features**
  Values replaced with actual scores rather than standardized values

# Substance Use Distributions

## Alcohol vs. Cocaine & Benzos

CL0 = Never Used

CL1 = Used over a Decade Ago

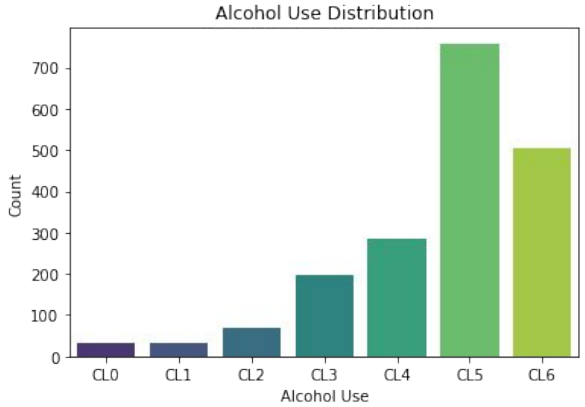
CL2 = Used in Last Decade

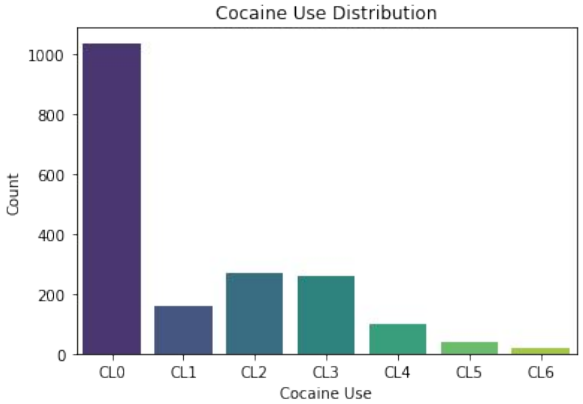
CL3 = Used in Last Year

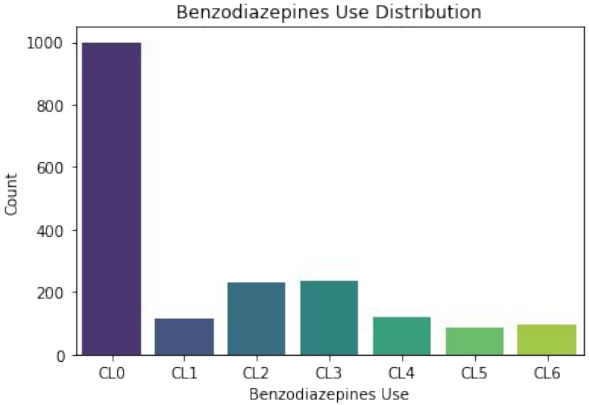
CL4 = Used in Last Month

CL5 = Used in Last Week

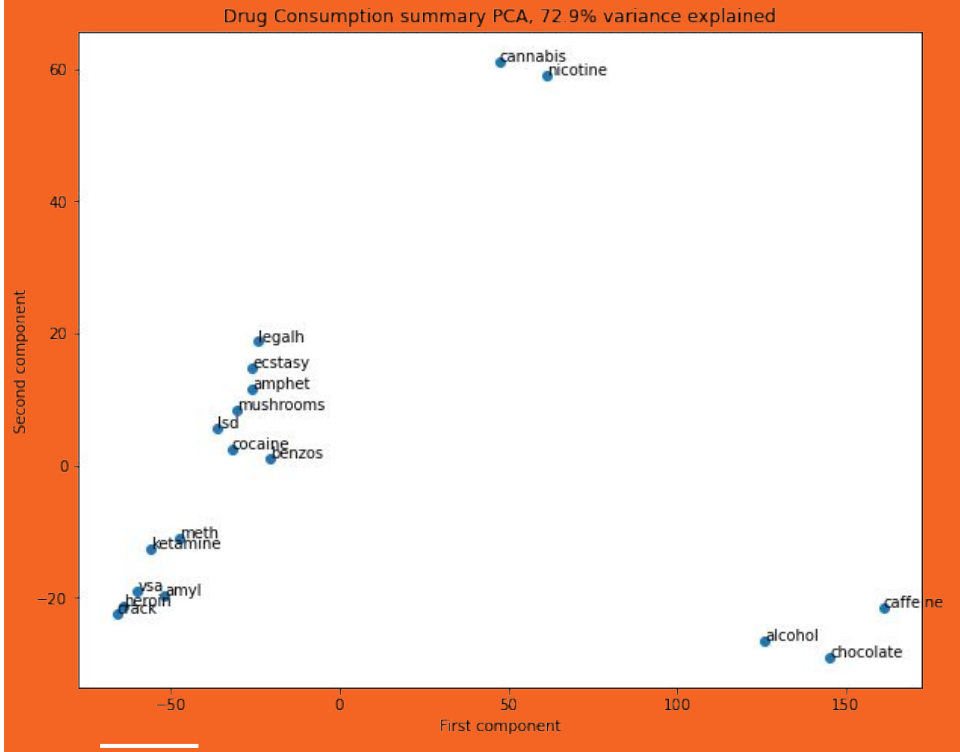
CL6 = Used in Last Day



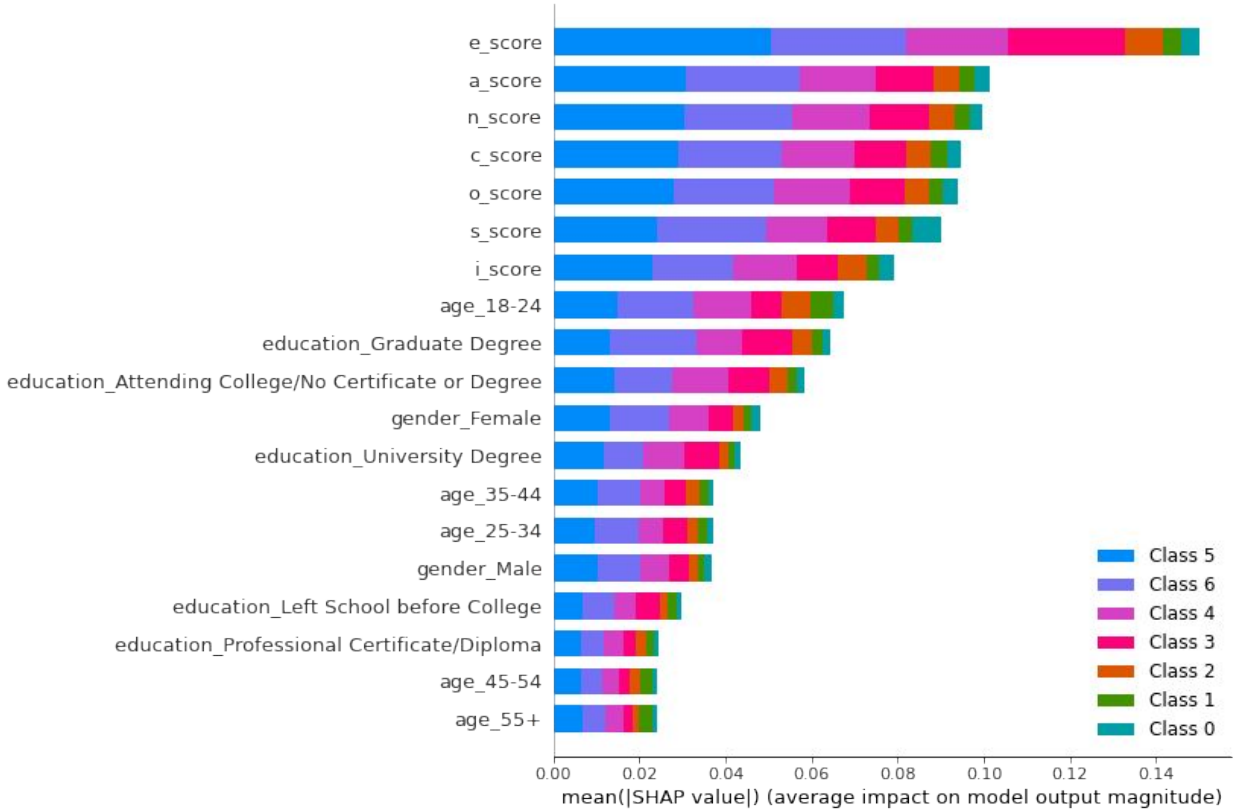



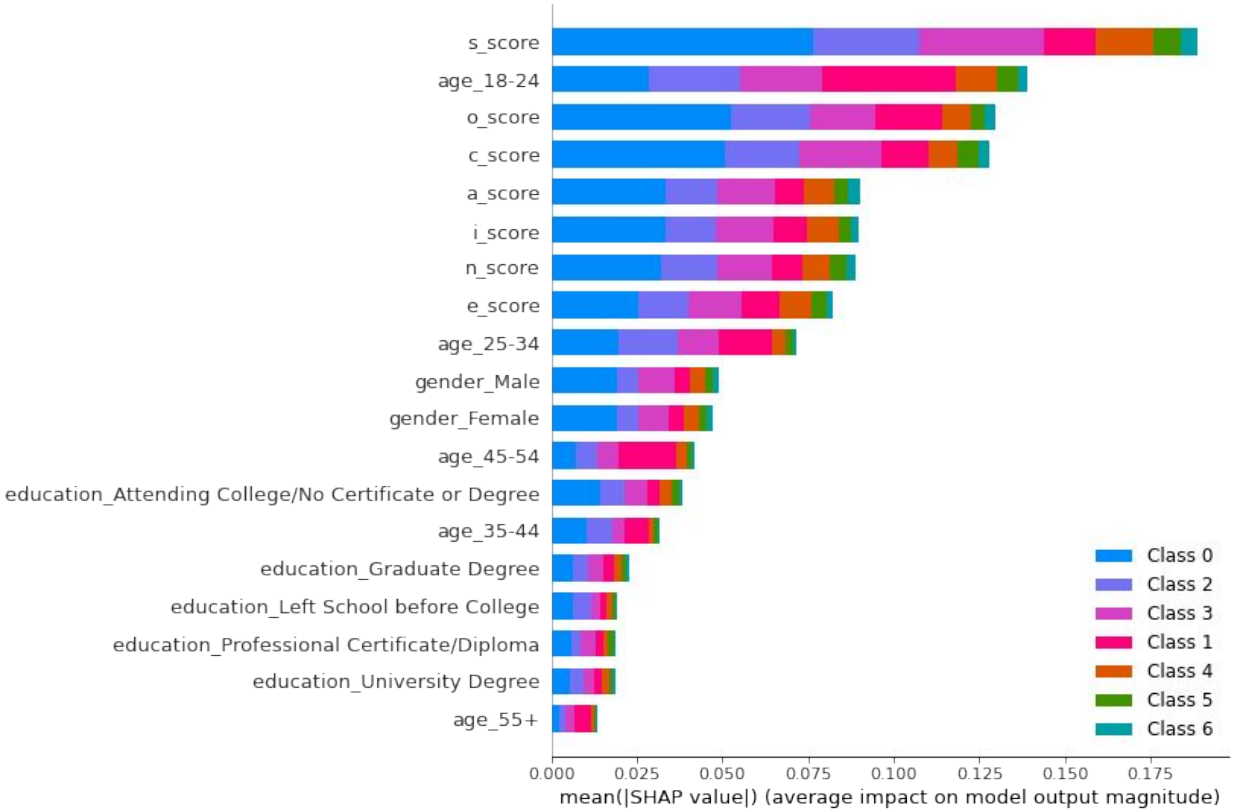
# Principal Components Analysis



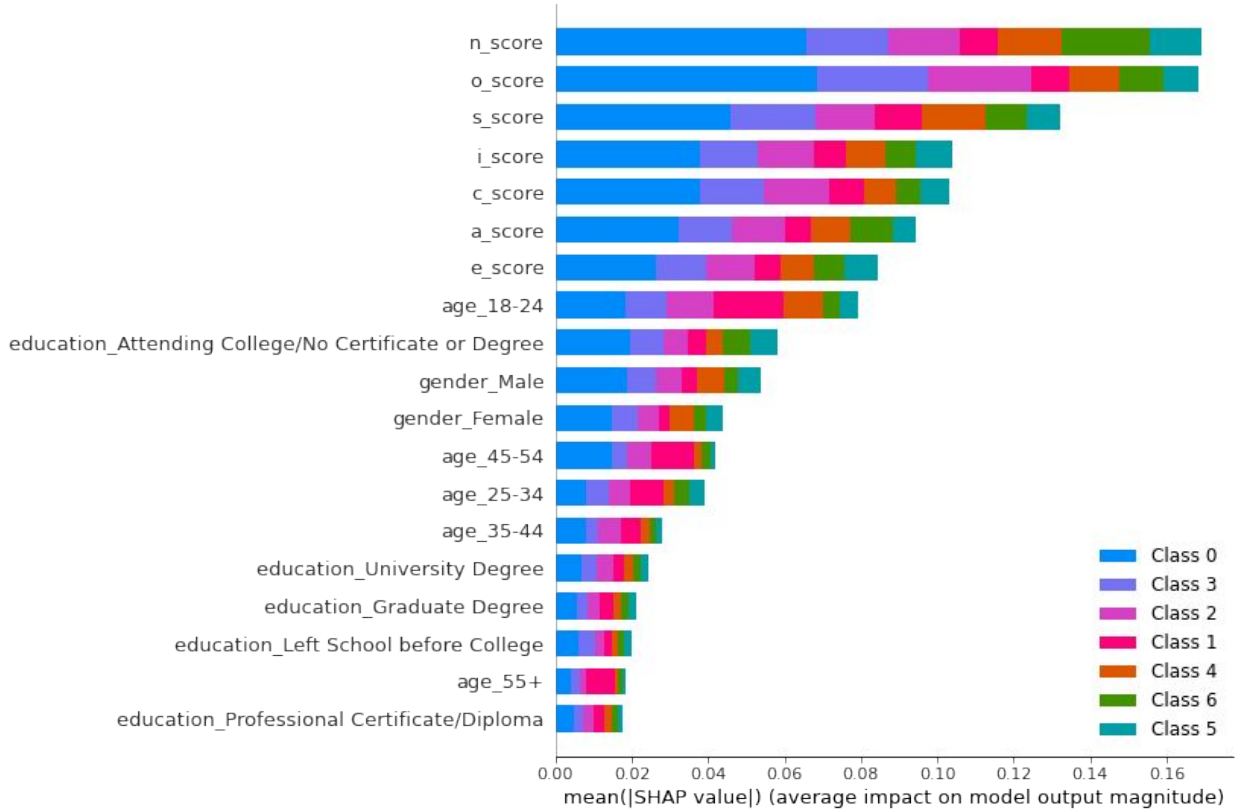
# Feature Importance: Alcohol



# Feature Importance: Cocaine



# Feature Importance: Benzodiazepines



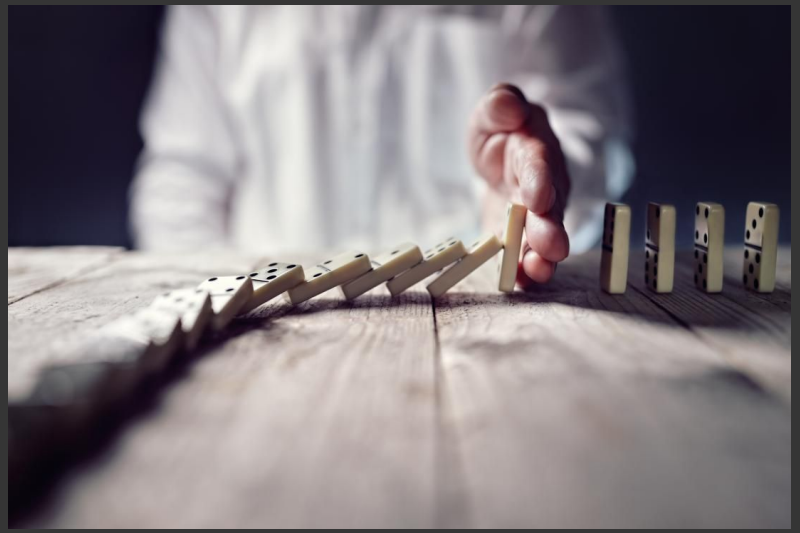
# Modeling

Target Variables: **Multi-Class** to **Binary**

Done **separately** for each substance

**75/25** Train - Test Split

**3 Classifiers Used:** Logistic Regression, Random Forest, K-Nearest Neighbors



# Procedure

Pick best model using **ROCAUC Score**

- Utilize **GridSearchCV** to tune hyperparameters

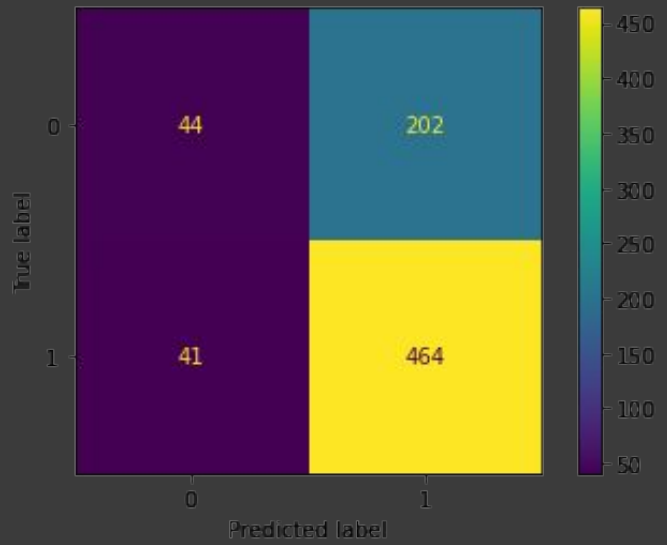
Find optimal Threshold using F beta score

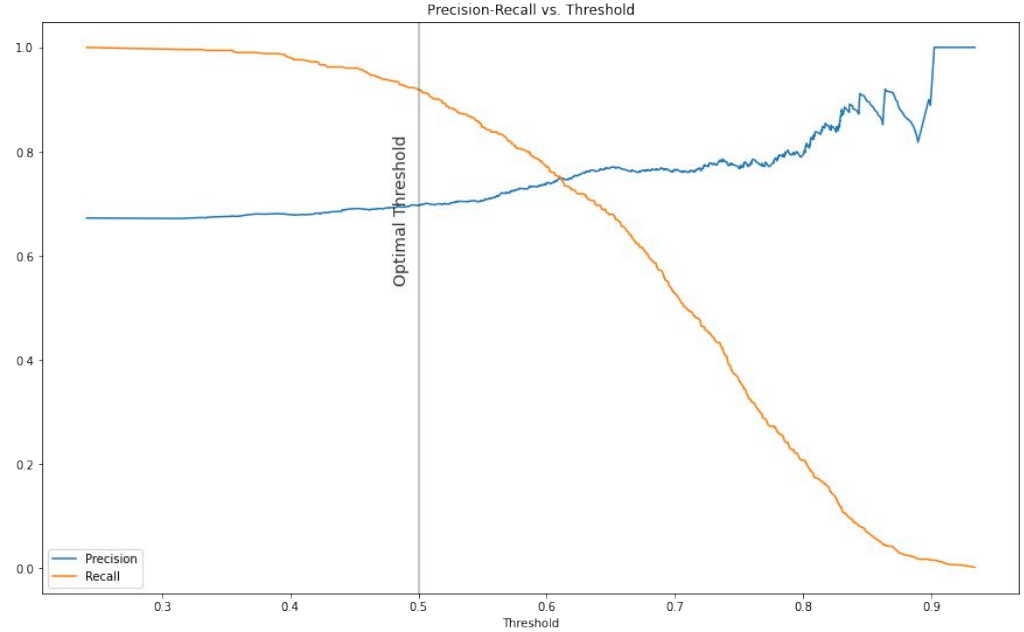
**REMEMBER!**

We would rather have false positives than false negatives - want to prioritize recall over precision

# Alcohol

Log. Regression  performs best

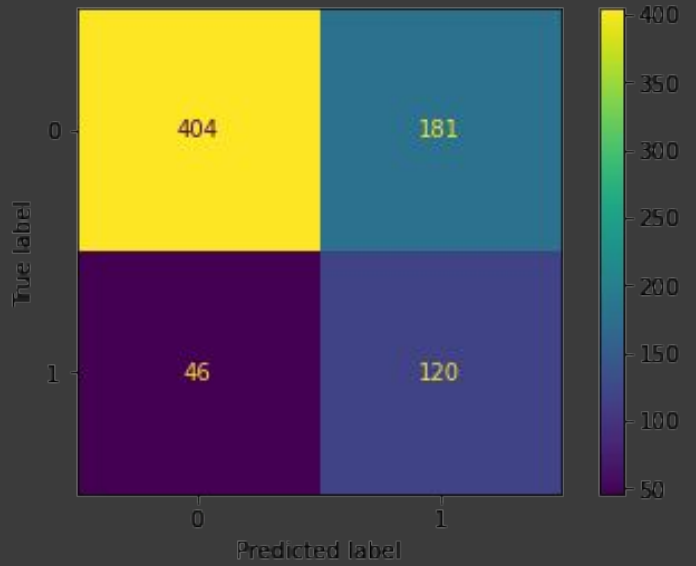


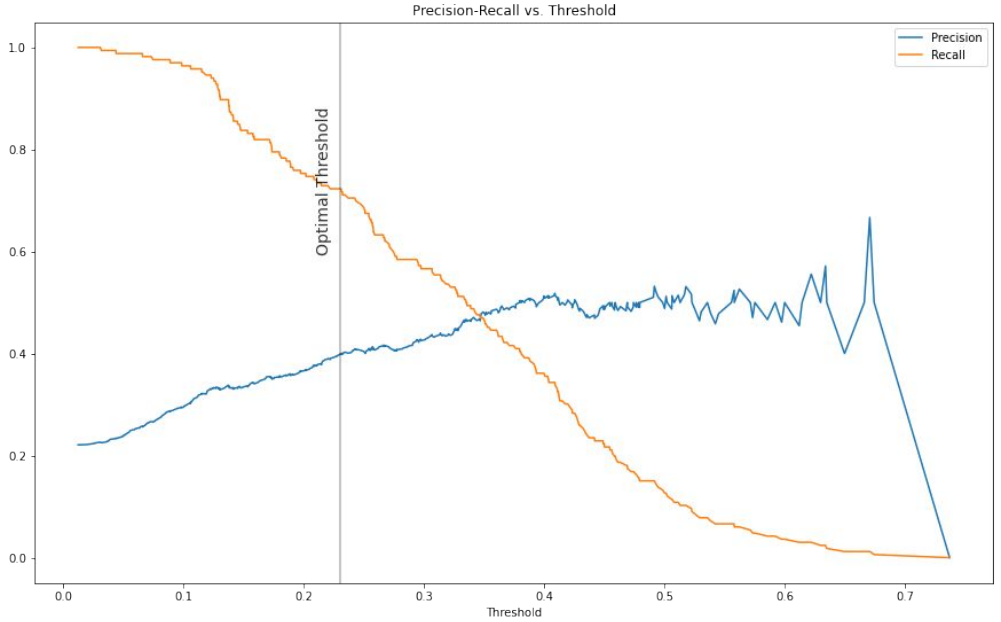




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.52 | 0.18 | 0.27 | 246 |
| 1 | 0.70 | 0.92 | 0.79 | 505 |
| accuracy | | | 0.68 | 751 |
| macro avg | 0.61 | 0.55 | 0.53 | 751 |
| weighted avg | 0.64 | 0.68 | 0.62 | 751 |

# Cocaine

Log. Regression performs best

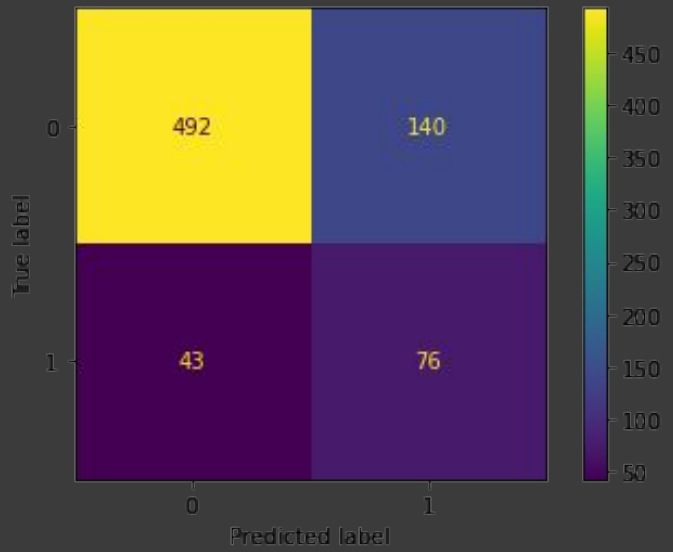


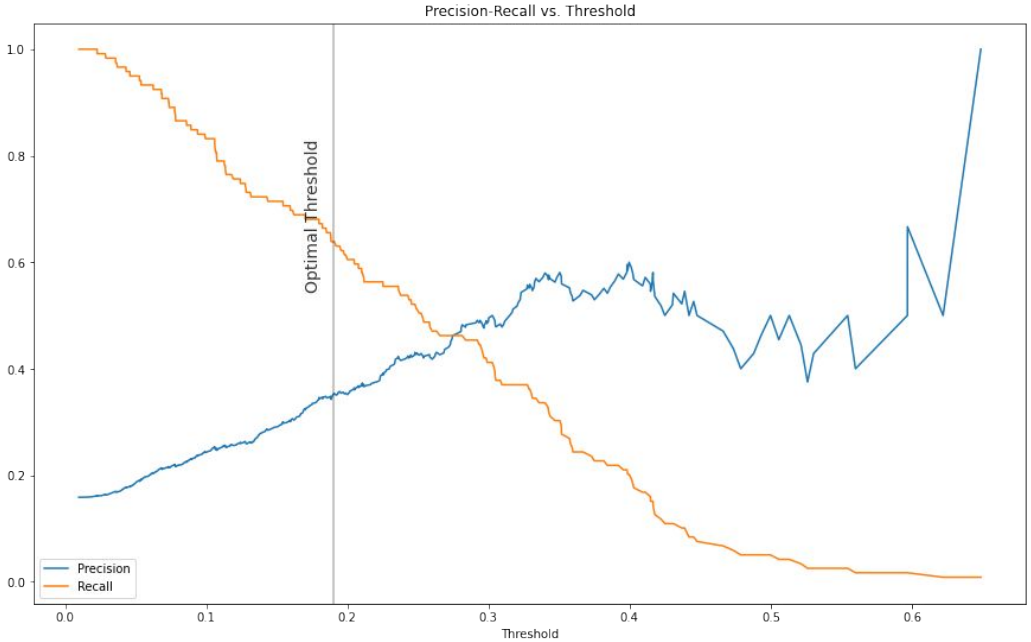




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.69 | 0.78 | 585 |
| 1 | 0.40 | 0.72 | 0.51 | 166 |
| accuracy | | | 0.70 | 751 |
| macro avg | 0.65 | 0.71 | 0.65 | 751 |
| weighted avg | 0.79 | 0.70 | 0.72 | 751 |

# Benzodiazepine

Log. Regression performs best



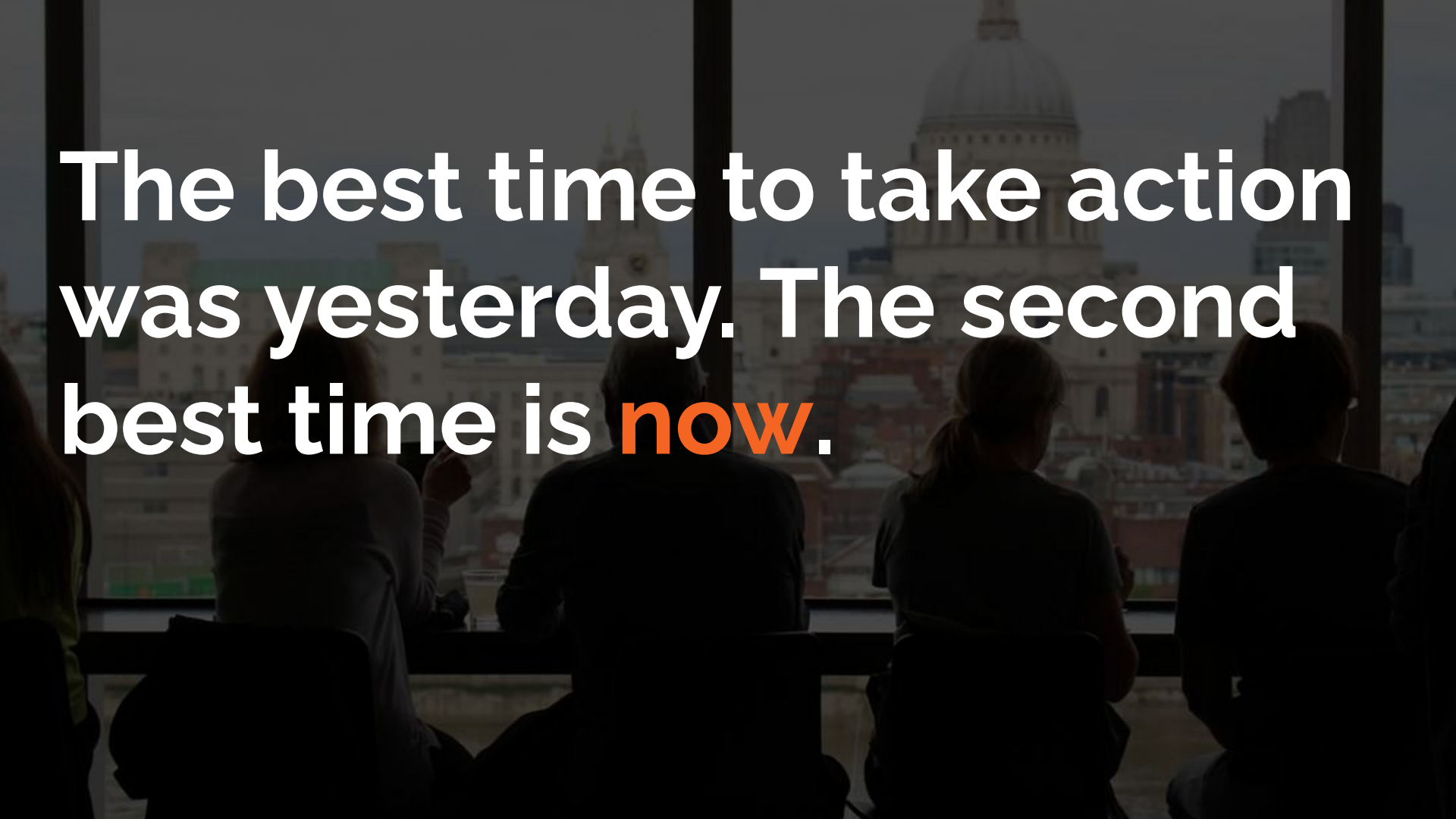

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.92 | 0.78 | 0.84 | 632 |
| 1 | 0.35 | 0.64 | 0.45 | 119 |
| accuracy | | | 0.76 | 751 |
| macro avg | 0.64 | 0.71 | 0.65 | 751 |
| weighted avg | 0.83 | 0.76 | 0.78 | 751 |

# Drawbacks

- **Lack of Features**
  Could help alcohol prediction model
- **Past Drug Use**
  Data only shows last use, not frequency

# Further Goals

- **Try More Classifiers**
  SVM, Naive Bayes, Ensembles, etc.
- **Redefine User vs. Non-User**
  Could be more or less strict in definition
- **Build Models for more Substances!**


The best time to take action was yesterday. The second best time is now.

# References

https://drugabusestatistics.org/

https://archive.ics.uci.edu/ml/machine-learning-databases/00373/drug_consumption.data