Capstone Two Presentation
"Personality Traits and Drug Consumption"

Thanh Nguyen

1

Introduction

- Understanding an individual's risk of drug consumption is critical
  - Drug overdose is widespread schools and colleges
  - Drug overdose deaths among seniors more than tripled in two decades
- Goal: to predict drug usage using demographic and personal factors
  - Amyl nitrite (Amyl thereafter)
  - Cannabis

2

Overview and agenda

- Overview of results
  - Amyl user prediction - low precision (50%)
  - Cannabis user prediction - high precision (90%)

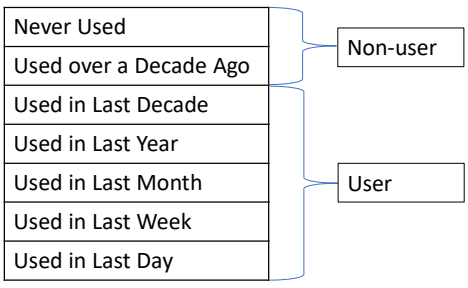
- Agenda
  - Dataset
  - Target, features, and correlation
  - Feature importance analysis
  - Modeling and performance on unseen data
  - Conclusion and future work

3

Dataset

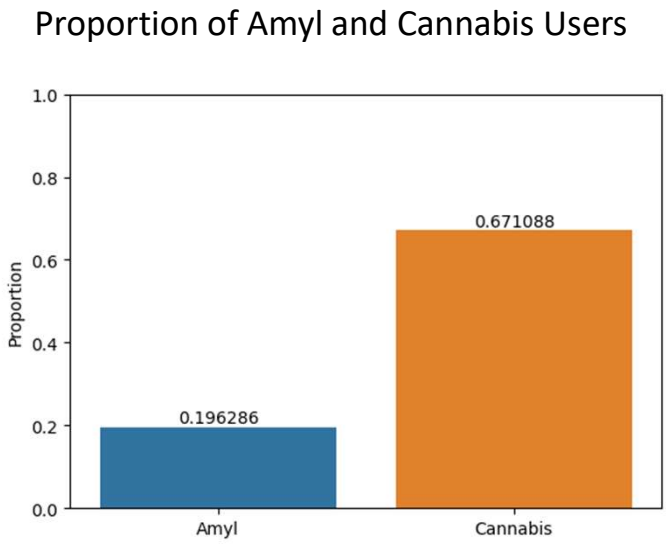
- University of California, Irvine's Machine Learning Repository
  - Online survey methodology
  - 1885 respondents
  - 18 legal and illegal drugs

- Target variable



4

Target



5

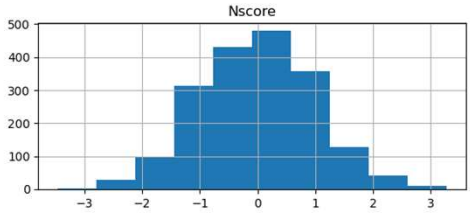
Features

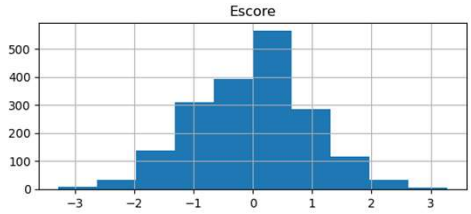
- Demographic factors
  - Gender
  - Age
  - Education
  - Country
- Personal traits
  - Big five personality traits
    - Neuroticism (measured by Nscore)
    - Extraversion (Escore)
    - Openness to experience (Oscore)
    - Agreeableness (Ascore)
    - Conscientiousness (Cscore)
  - Impulsivity and Sensation Seeking (SS)

6

Features

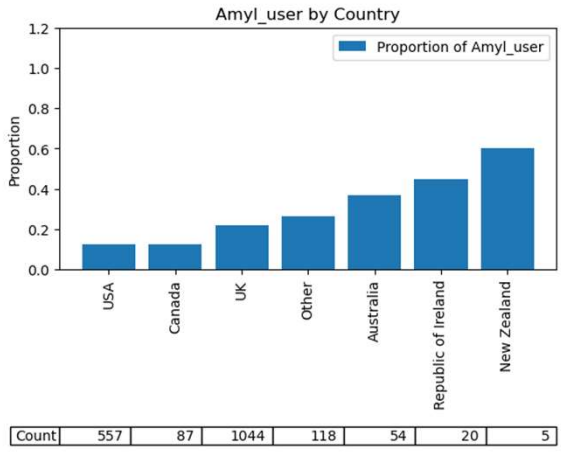
- Big five personality traits' measurement
  - For each trait, 12 questions asked
  - Answers: 0 = 'Strongly Disagree', 1 = 'Disagree', 2 = 'Neutral', 3 = 'Agree', 4 = 'Strongly Agree')
  - Summarizing and standardizing

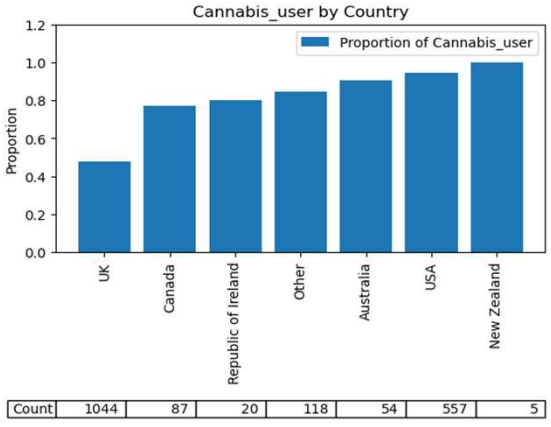




7

Target-feature relationship

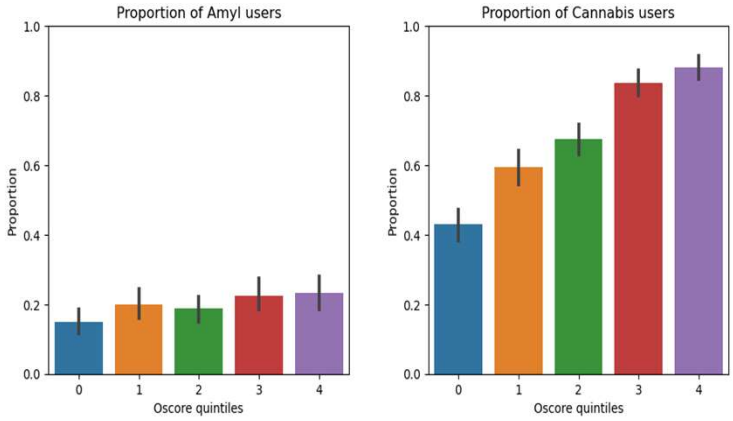




8

Target-feature relationship

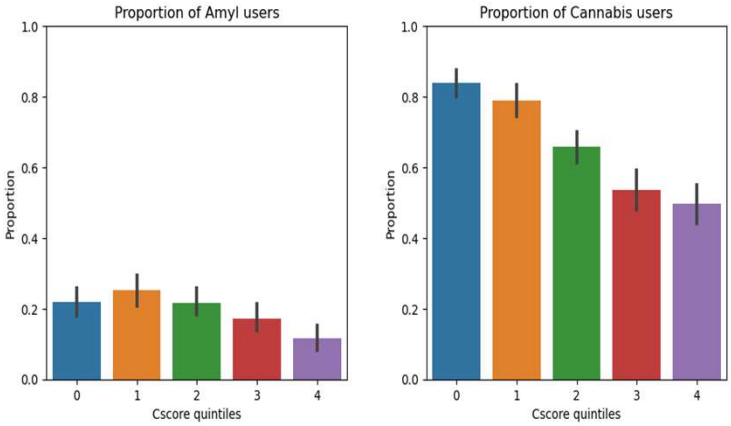
Drug Consumption Across Oscore (Openness to Experience) Quintiles



9

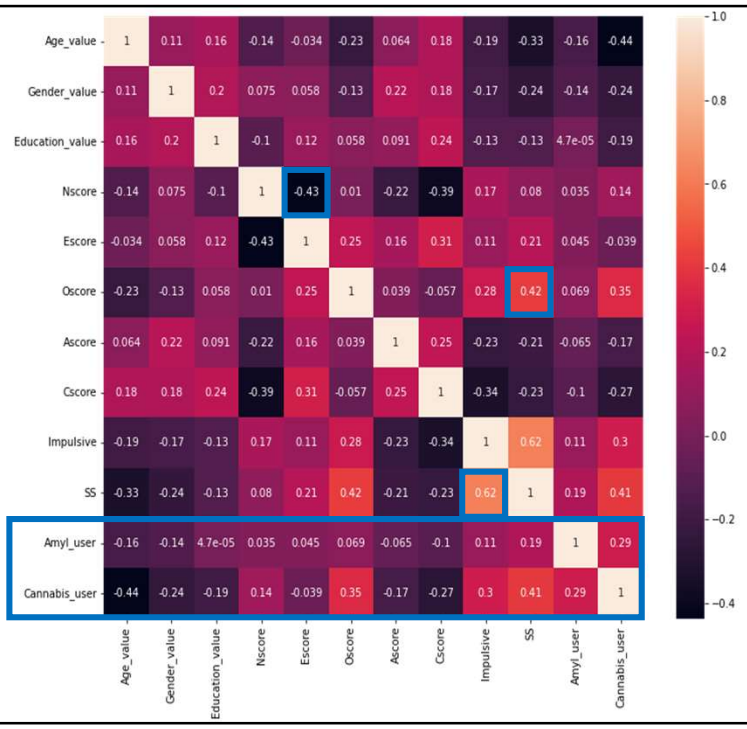
Target-feature relationship

Drug Consumption Across Cscore (Conscientiousness) Quintiles



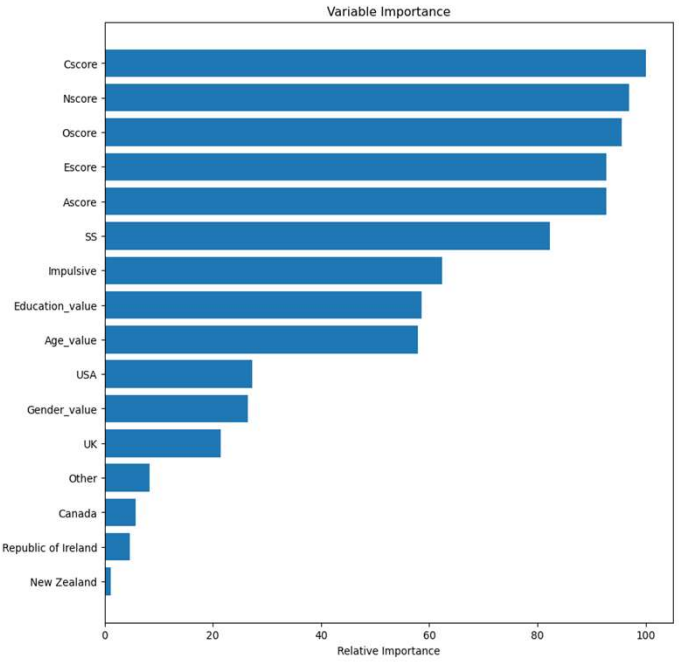
10

Correlation Matrix



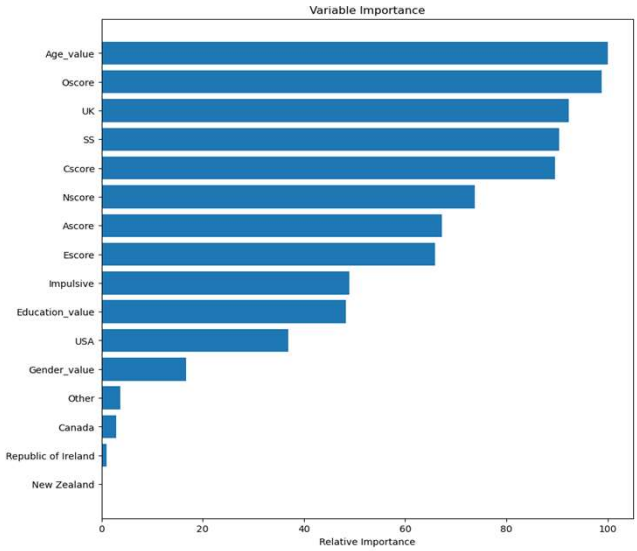
11

Relative Feature Importance in Predicting Amyl Consumption



12

Relative Feature Importance in Predicting Cannabis Consumption



13

Modeling

- Hyperparameter tuning
  - Whole samples split into Train set (80%) and Test set (20%)
- Choosing optimal classification threshold
  - Train set split into Main set (70%) and Validation set (30%)
- Evaluating performance on Test set

14

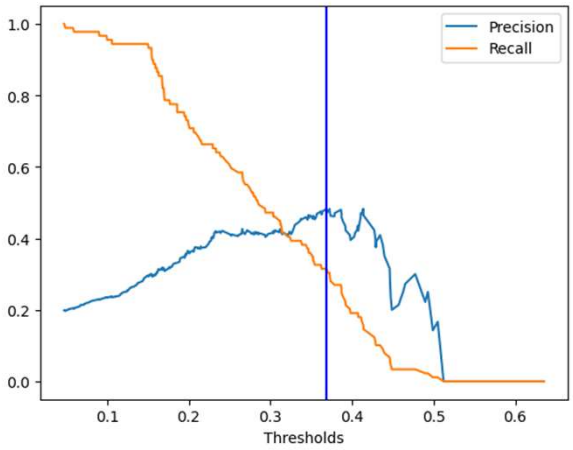
Hyperparameter Tuning for Amyl Users Prediction

Model Name	Best Parameters	Optimal Feature Set	Cross-Validation Score (roc_auc)
Random Forest	max_depth = 5 max_features = 'auto' n_estimators = 50	All features.	0.7533
Gradient Boosting	learning_rate = 0.05 max_depth = 3 max_features = 8 min_samples_split = 2 n_estimators = 50 subsample = 1	All features.	0.7571
Logistic regression	C = 1 penalty = 'l1' solver = 'liblinear'	All features.	0.7486

15

Optimal Threshold and Test Performance

Precision and Recall Curves for Amyl Users Prediction



Test Performance

Precision	Recall	Accuracy
50%	23%	80.4%

16

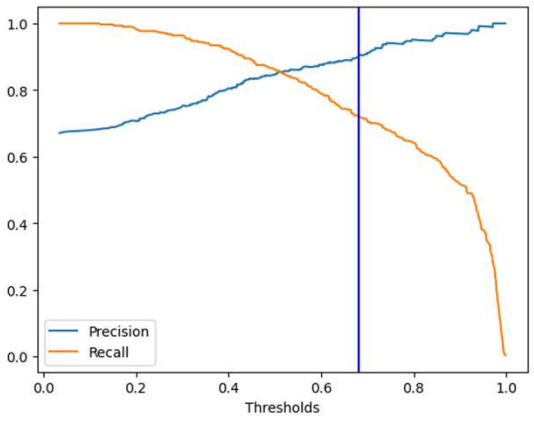
Hyperparameter Tuning for Cannabis Users Prediction

Model Name	Best Parameters	Optimal Feature Set	Cross-Validation Score (roc_auc)
Random Forest	max_depth = 5 max_features = 0.3 n_estimators = 50	All features.	0.8778
Gradient Boosting	learning_rate = 0.02 max_depth = 3 max_features = 6 min_samples_split = 6 n_estimators = 100 subsample = 1.0	All features.	0.8780
Logistic regression	C = 0.1 penalty = 'l2' solver = 'sag'	All features.	0.8781

17

Optimal Threshold and Test Performance

Precision and Recall Curves for Cannabis Users Prediction



Test Performance

Precision	Recall	Accuracy
90.3%	77.1%	79%

18

Conclusion

- Amyl user prediction has low precision (50%)
  - Use as secondary tool in the toolbox of school and healthcare professionals
- Cannabis user prediction has high precision (90%)
  - Use as primary tool

19

Future Work and References

- Data validity check
- Interaction terms among demographic and personal features
- References:
  - E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan and A. N. Gorban, "The Five Factor Model of personality and evaluation of drug consumption risk.," arXiv [Web Link], 2015
  - UCI Machine Learning Repository: Drug consumption (quantified) Data Set
  - Opioids: Drug overdose deaths among seniors have more than tripled in 2 decades (cnbc.com)

20