Illicit Drug Use Predictor

Ake Paramadilok
Metis chi20_ds13
2020.02.12

Did you know?

More than 700,000 Americans died from drug overdose between 1999 and 2017

Did you know?

More than **21 million** Americans have an **addiction**...

Did you know?

… only **10%** are **treated**.

Mission

Predict the likelihood that an individual is a user of illicit drugs based on their personality traits and demographics.

Purpose

Counseling|therapy

- ID patients at high risk
- Provide early interventions
- Advise treatment plan

Data

1882 samples

Data

1882 samples

Demographics

- Age
- Gender
- Country
- Education

Data

1882 samples

Demographics

- Age
- Gender
- Country
- Education

Personality Score

- Neuroticism
- Extrovertism
- Openness to Experiences
- Agreeableness
- Conscientiousness
- Impulsivity
- Sensation Seeking

Data

1882 samples

Demographics

- Age
- Gender
- Country
- Education

Personality Score

- Neuroticism
- Extrovertism
- Openness to Experiences
- Agreeableness
- Conscientiousness
- Impulsivity
- Sensation Seeking

Drug Use

- Chocolate
- Coffee
- Alcohol
- Nicotine
- Cannabis
- Cocaine
- Crack
- Ecstacy
- Heroine
- Ketamine
- Legal Highs
- LSD
- Meth
- Mushrooms
- VSA

Defining a ‘User’

USER

Defining a ‘User’

USER

Illicit Drugs

- Chocolate
- Coffee
- Alcohol
- Nicotine
- Cannabis
- Cocaine
- Crack
- Ecstacy
- Heroine
- Ketamine
- Legal Highs
- LSD
- Meth
- Mushrooms
- VSA

Defining a ‘User’

USER

Illicit Drugs

- Chocolate
- Coffee
- Alcohol
- Nicotine
- **Cannabis**
- **Cocaine**
- **Crack**
- **Ecstacy**
- **Heroine**
- **Ketamine**
- **Legal Highs**
- **LSD**
- **Meth**
- **Mushrooms**
- **VSA**

Defining a ‘User’

USER

Usage

- Never
- ≥ 10 years ago
- 1 - 9 years ago
- This year
- This month
- This week
- Yesterday

Illicit Drugs

- Chocolate
- Coffee
- Alcohol
- Nicotine
- **Cannabis**
- **Cocaine**
- **Crack**
- **Ecstacy**
- **Heroine**
- **Ketamine**
- **Legal Highs**
- **LSD**
- **Meth**
- **Mushrooms**
- **VSA**

Defining a ‘User’

USER

Usage

- Never
- ≥ 10 years ago
- 1 - 9 years ago
- **This year**
- **This month**
- **This week**
- **Yesterday**

Illicit Drugs

- Chocolate
- Coffee
- Alcohol
- Nicotine
- **Cannabis**
- **Cocaine**
- **Crack**
- **Ecstacy**
- **Heroine**
- **Ketamine**
- **Legal Highs**
- **LSD**
- **Meth**
- **Mushrooms**
- **VSA**

Defining a ‘User’

USER

55% of data

Usage

- Never
- ≥ 10 years ago
- 1 - 9 years ago
- **This year**
- **This month**
- **This week**
- **Yesterday**

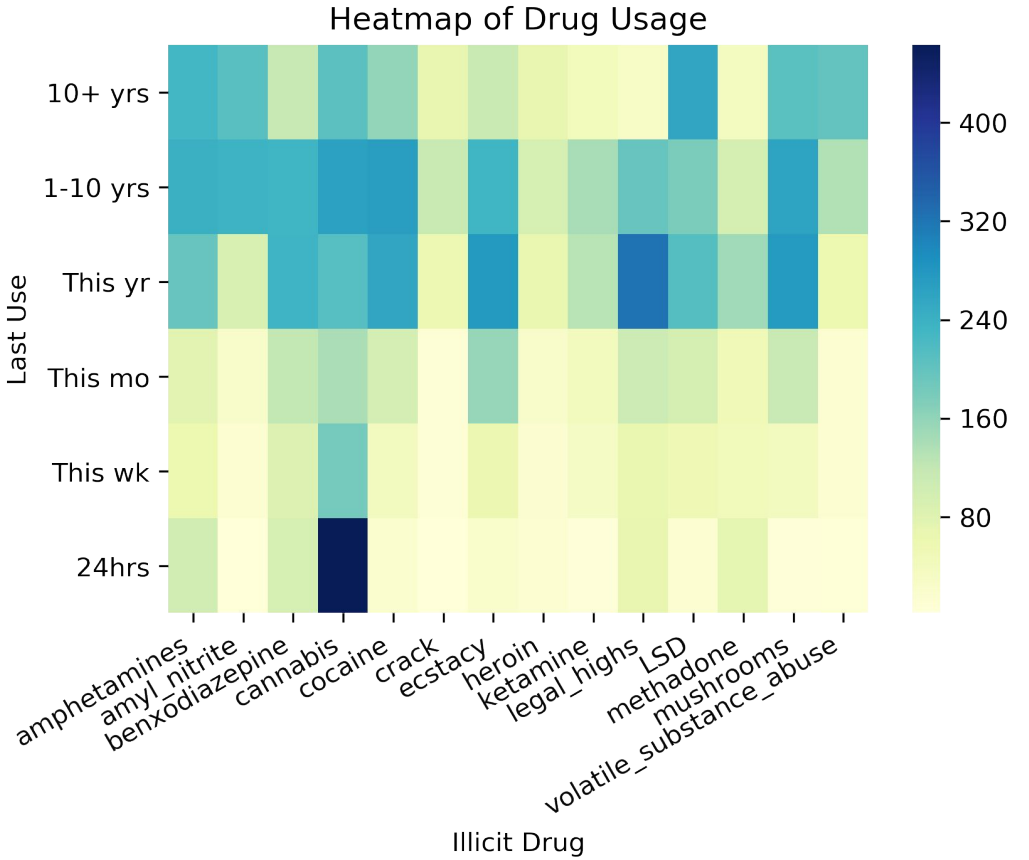
Illicit Drugs

- Chocolate
- Coffee
- Alcohol
- Nicotine
- **Cannabis**
- **Cocaine**
- **Crack**
- **Ecstacy**
- **Heroine**
- **Ketamine**
- **Legal Highs**
- **LSD**
- **Meth**
- **Mushrooms**
- **VSA**

Defining a ‘User’

USER

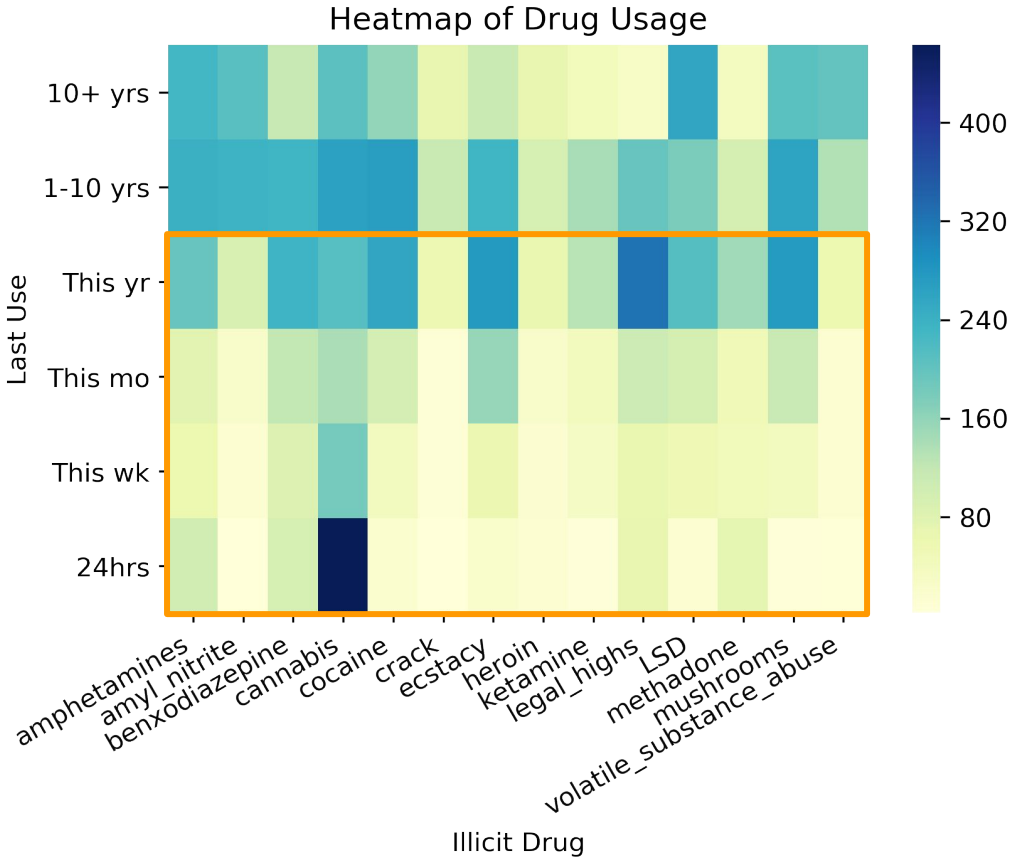
55% of data



Defining a ‘User’

USER

55% of data



Parameters

Feature engineering driven by clinical research on personality traits and drug use.[1]

- High N, E and low C are highly correlated with hazardous health behaviours.
- Low C and high socioeconomic status correlated with illicit drug use.
- High N and low A and C associated with higher risk of drug use.
- Increasing N and O increases risk of drug use
- Increasing C and A decreases risk of drug use.
- Sensation seeking is high for recreational drug users.
-

External data sources scraped by country

- GDP Per Capita
- Sex specific life expectancy
- Mean education level
- Expected education level
- Divorce rate

1. arXiv:1506.06297

Classification Model

ID and help as many ‘users’ as possible

Classification Model

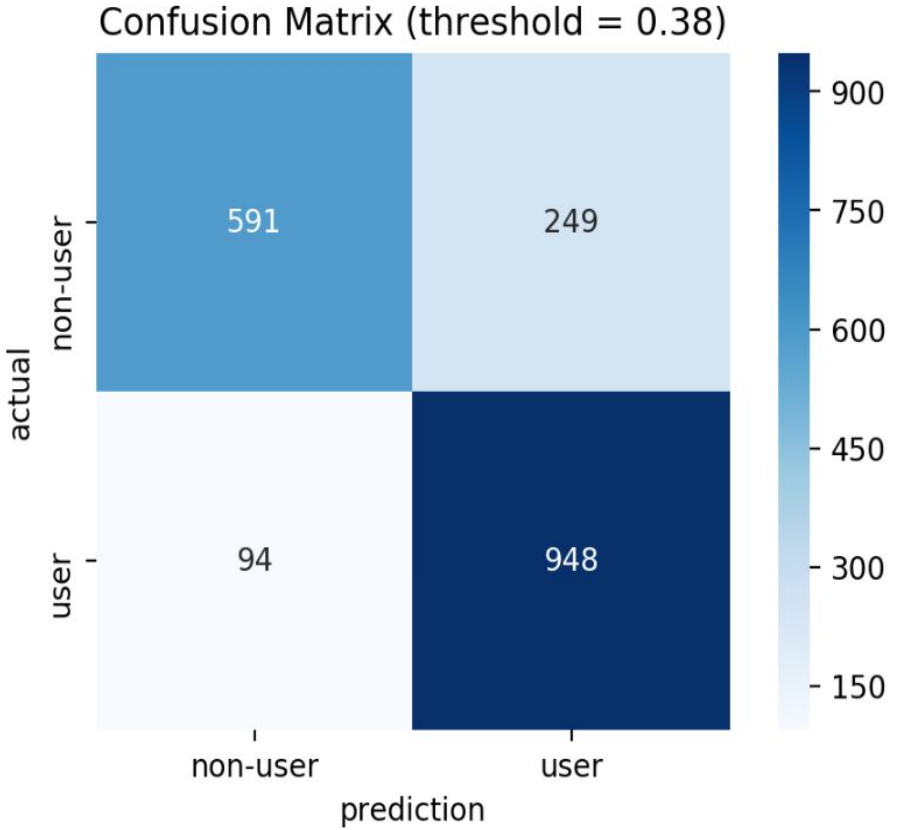
RECALL

Best Model: Random Forest

RECALL

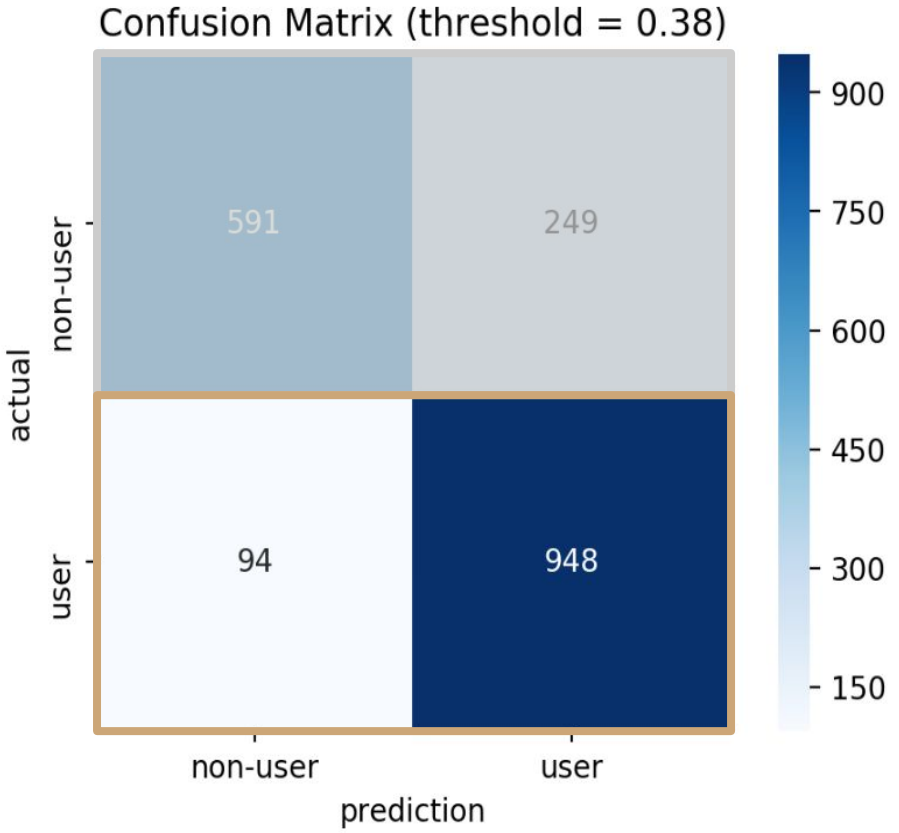
Random Forest

RECALL



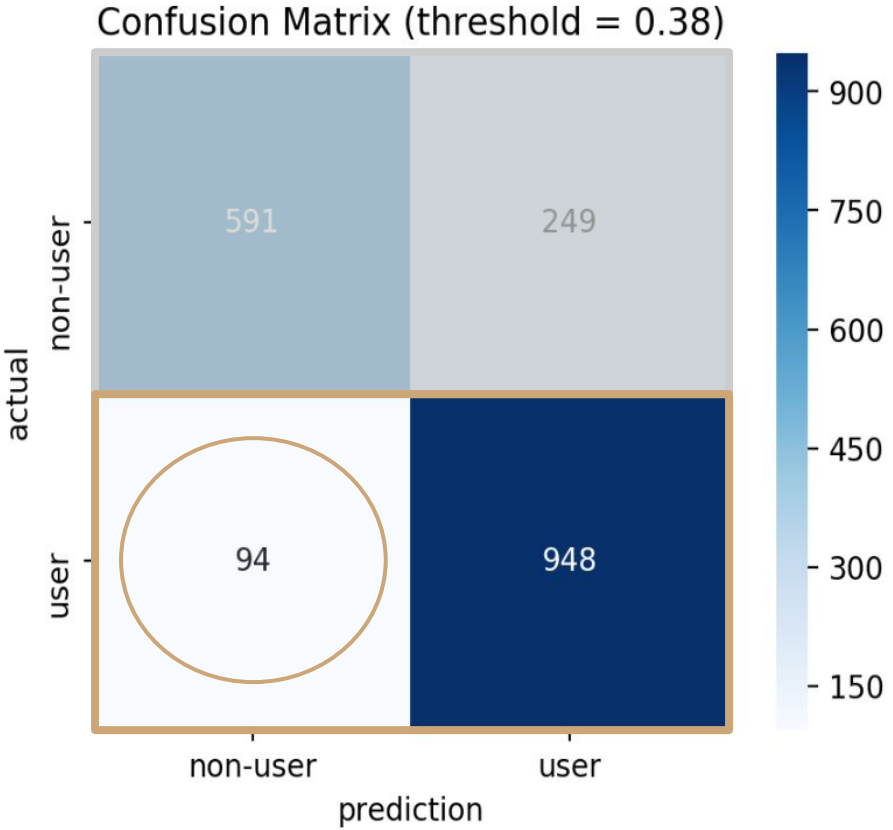
Random Forest

RECALL



Random Forest

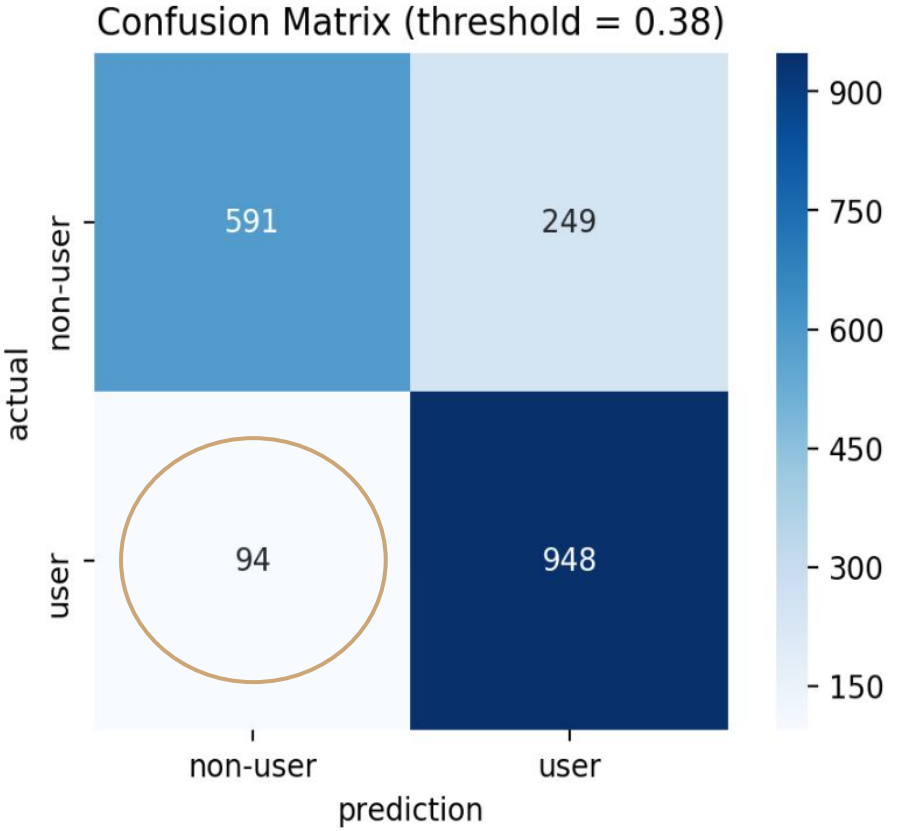
RECALL



Random Forest

>90%

RECALL

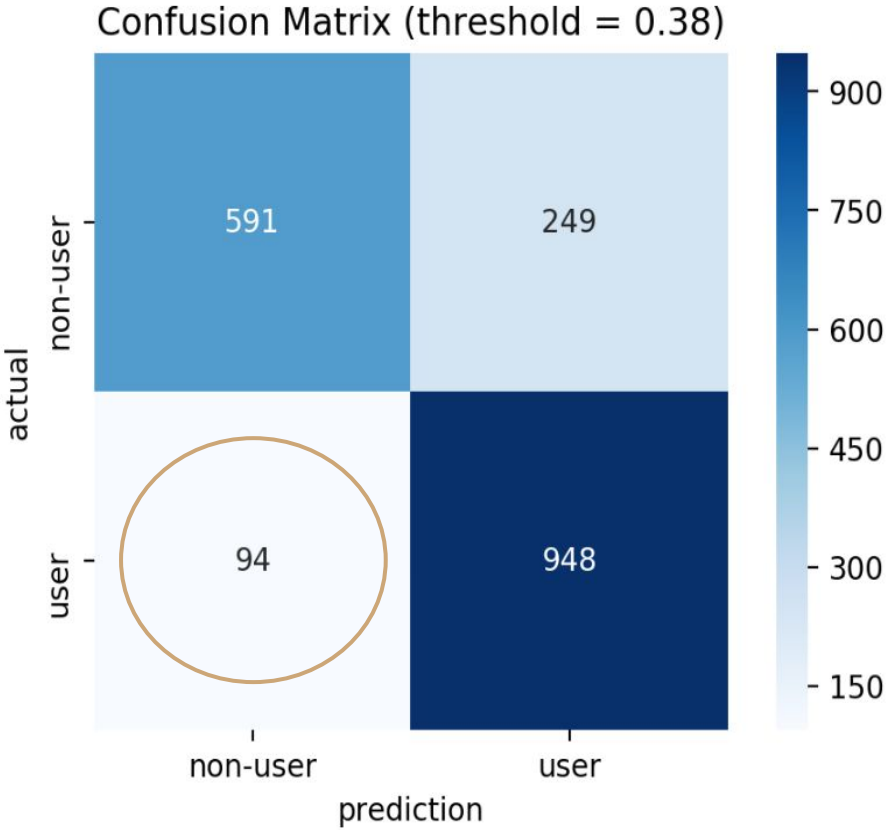


Random Forest Performance

>90%

RECALL

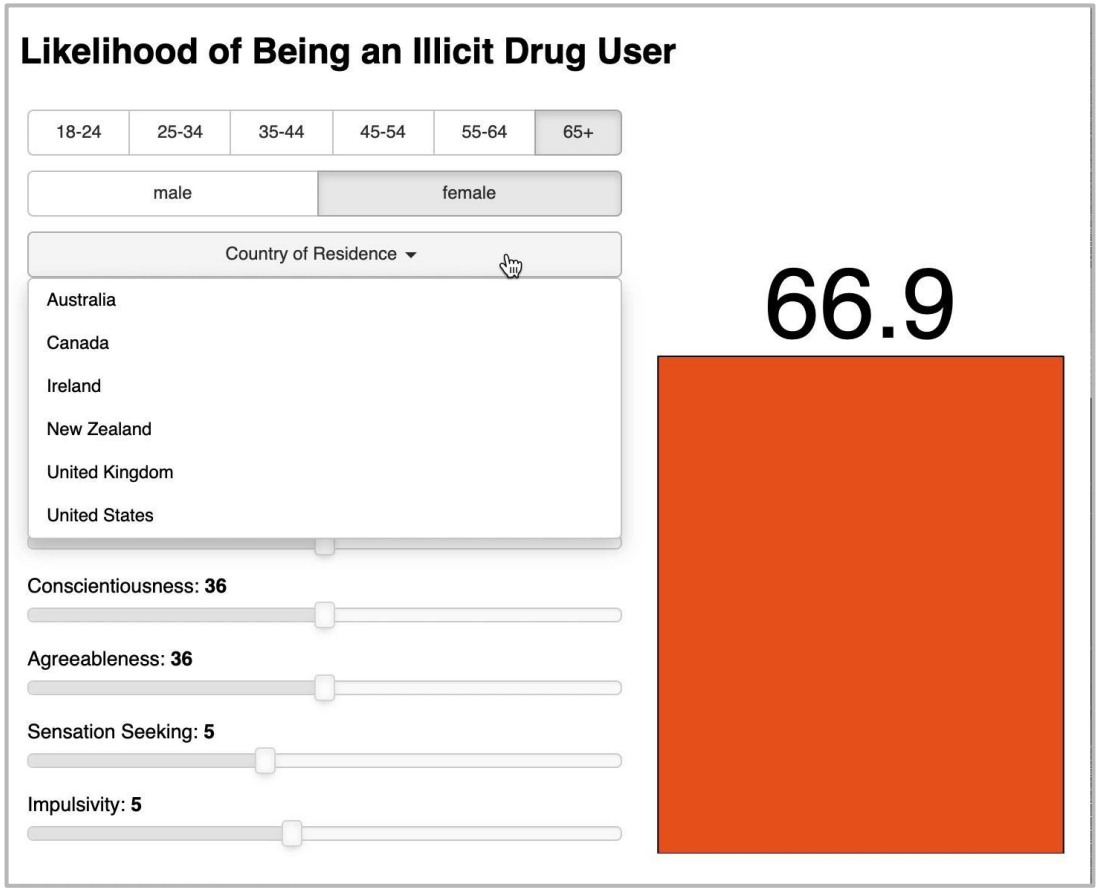
Model identifies 9 out of 10 illicit drug users



Feature	Importance
$\frac{GDP}{c}$	0.255
$\frac{(G + 1)^2 (s + o + i)}{(c + A)}$	0.190
sensation seeking	0.124
$((n/n\_ave) * e * 2 * s) - 85 * (c/c\_ave)^2$	0.11
$\frac{education}{expected\ education}$	0.089

GDP: gross domestic product | G: gender | c: conscientiousness | s: sensation seeking | o: openness | i: impulsivity

Prediction Tool: Based on Random Forest Model



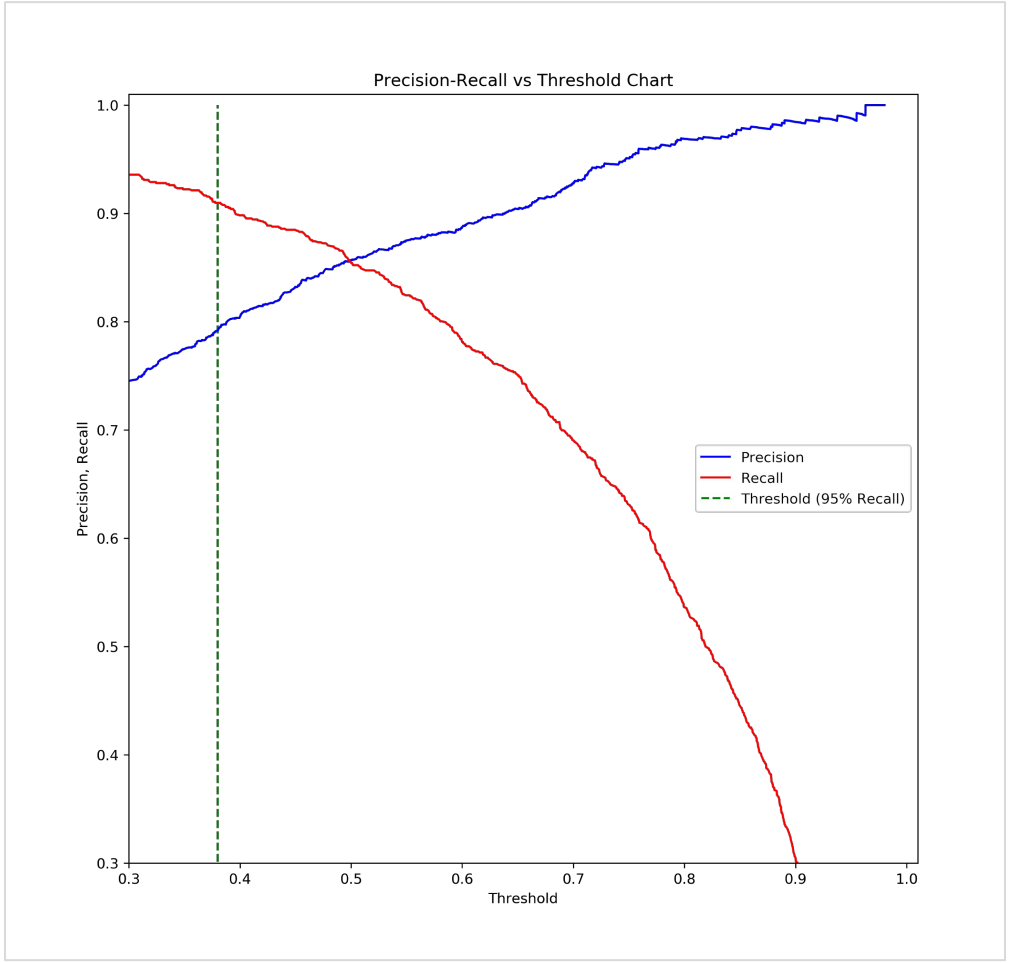
Key Conclusions

- Model prioritizes identifying users over accuracy of identification
- Recall is ~90% for Random Forest classification model
- GDP and conscientiousness are highly impactful to the model
- Gender, education and personality traits contribute to engineered features
- Predictor tool can be used to inform early intervention and treatment focus

Questions?

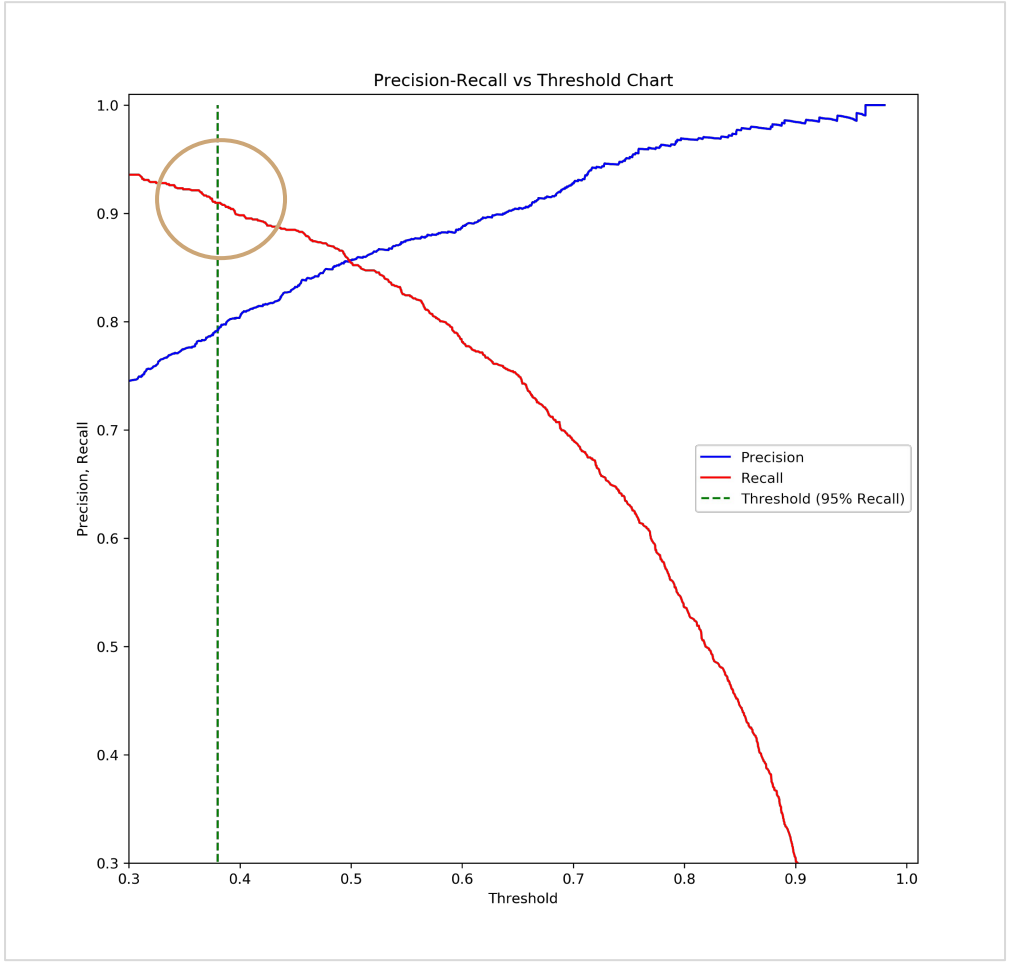
Prioritizing Recall

Reaching as many ‘users’ as possible

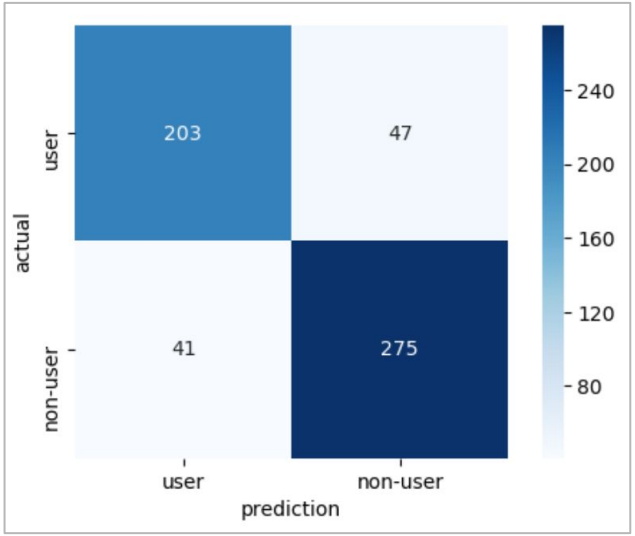


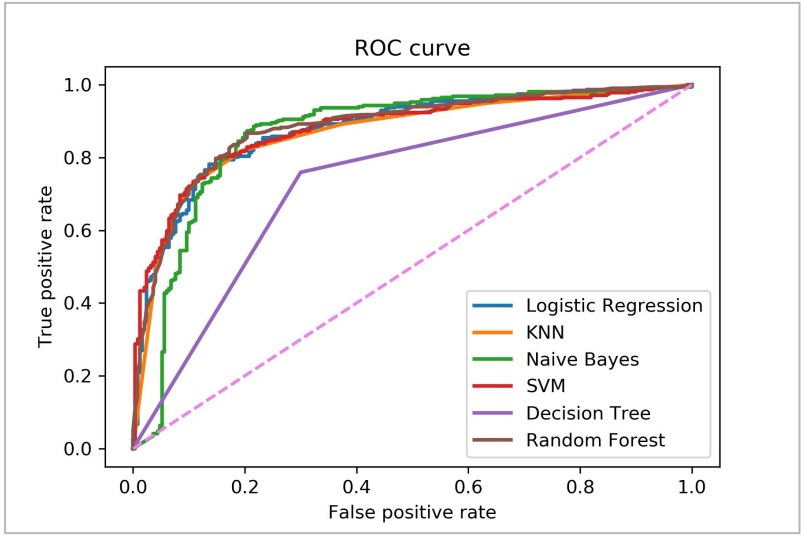
Prioritizing Recall

Reaching as many ‘users’ as possible



MVP: Illicit Drug User?



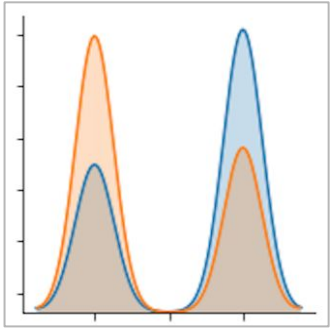


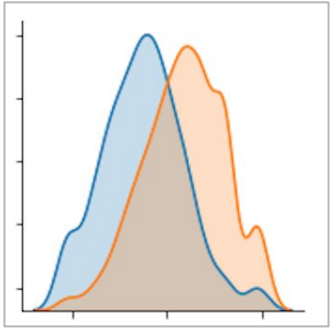
Baseline: 55.83%

Random Forest: 79.61%

Good feature separation

Next: Tuning/feature engineering





	Logistic Regression	KNN	Naive Bayes	SVM	Decision Tree	Random Forest
Accuracy(train)	0.7908	0.7529	0.7529	0.7680	0.7165	0.7961
Precision	0.8276	0.7958	0.8666	0.8180	0.7444	0.8265
Recall	0.7860	0.7449	0.6530	0.7476	0.7394	0.7956
F1	0.8055	0.7688	0.7445	0.7806	0.7478	0.8136
AUC	0.8789	0.8692	0.8652	0.8797	0.7306	0.8868