Project 4 Drugs

by Kal Lemma

sponsored by D.A.R.E

Questions and Approach

Drugs Consumption Data Set:

- 1,885 records
  - 12 Personal Measurement attributes (exs: scores on extraversion, impulsiveness)
  - Usage activity of 18 different Legal and Illegal Drugs (exs: chocolate, alcohol, weed, heroin)
    - Based on 7 classes of Never Used, Used in the Last Decade, Year, Month, Week, Day

1. What should I look for? - Strong Drugs
2. Can I answer it? - Depends on what's ‘considered’ Strong Drugs
3. Start EDA and Cleaning.

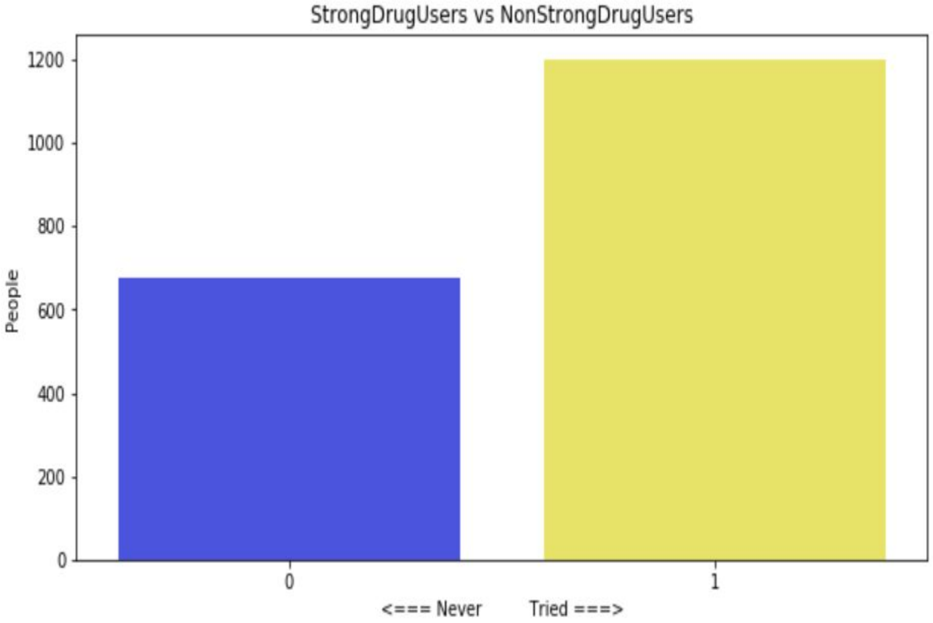
Exploratory Data Analysis/ Test Variable

1. Assigning ‘Labels’, turning values numeric, had no Missing values
2. Removing Bias from liars in the ‘Semer’ column (Fake Drug*)
3. Cleaned and created dummy variables for Gender and Lived in the U.K.
    a. The United Kingdom made up 55% of dataset

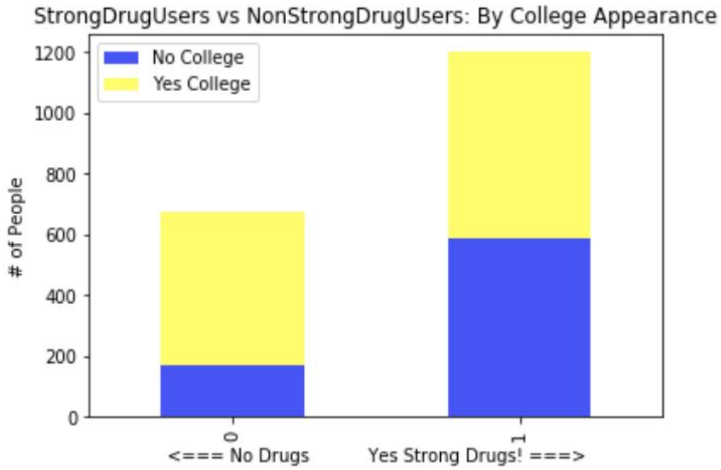
Test Variable:

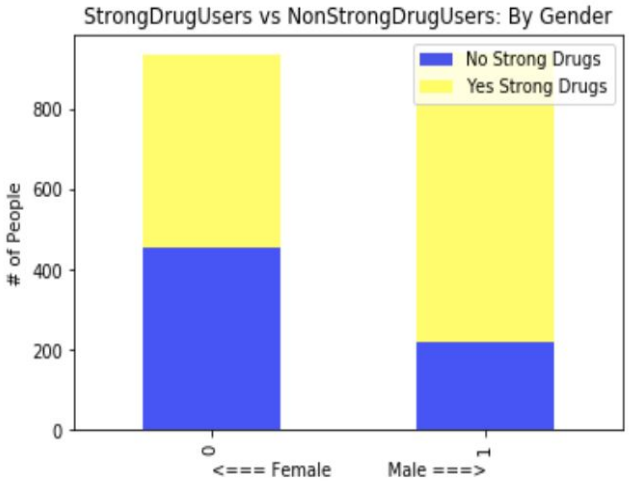
- **Strong Drugs**: Cocaine, Crack, Ecstasy, Heroin, Ketamine, LSD, Mushrooms
    - Had to clean values, summed hard drugs use, turned values binary
        - 0 for never Used, 1 for used before

So c’mon, who does it?



```
HardDrugUse
HardDrugUse
0      677
1     1200
```



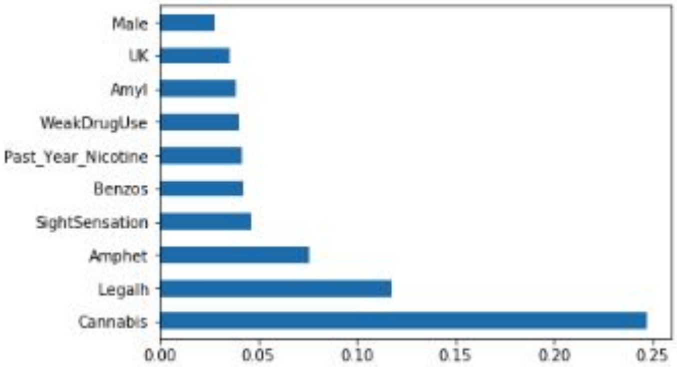


Feature Engineering and Selection

Indicator / Interaction features

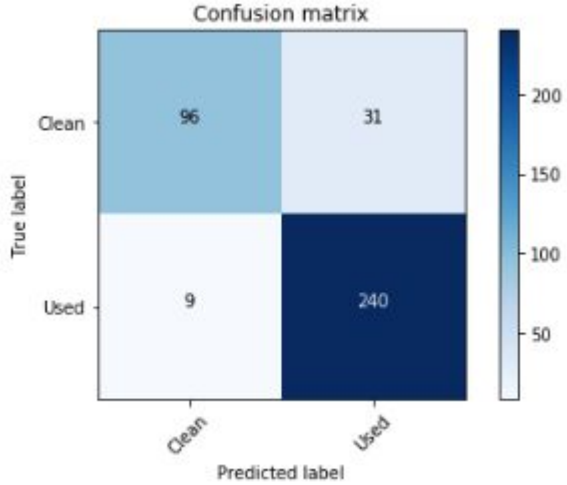
1. Age: Classified into three categories of Young, Middle-Aged, Old
   a. Dummy's, dropping old
2. Education: had multiple selections for years attained, chose baseline column for whether they made it to College, indicator variable
3. Nicotine: indicator whether person smoked in past year
4. WeakDrugUse: Interaction feature summing use for legal 'weaker' drugs

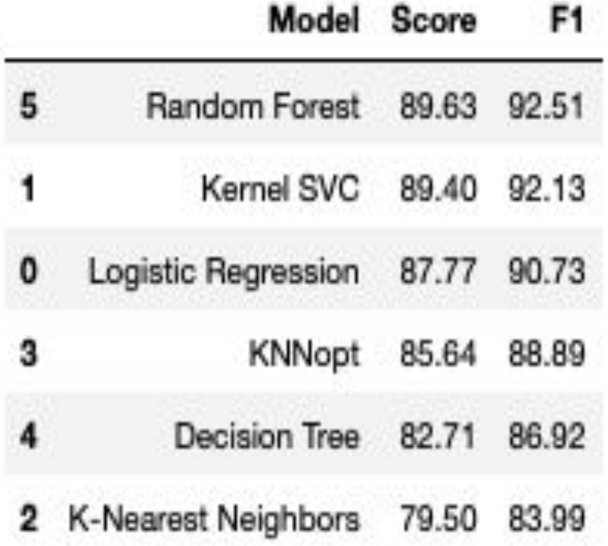
Feature Importance for top 10 performers;



Models

1. Good old Logistic Regression performed well but the Support Vector Classifier one upped
2. Random Forest Outshines, cm below





	Model	Score	F1
5	Random Forest	89.63	92.51
1	Kernel SVC	89.40	92.13
0	Logistic Regression	87.77	90.73
3	KNNopt	85.64	88.89
4	Decision Tree	82.71	86.92
2	K-Nearest Neighbors	79.50	83.99

Sum Up:

For my top performing model, using Random Forest and gridsearch for tuning Hyper-Parameters, I am able to classify if a person has done any of the ‘Strong-Drugs’ in their life, given my feature input data, at a 92% F1 score.

Data would suggest more people experiment than what may be commonly thought.