OKCupid Date-A-Scientist

Erik Owens
Codecademy
Portfolio Project
January 2021

Project Scope

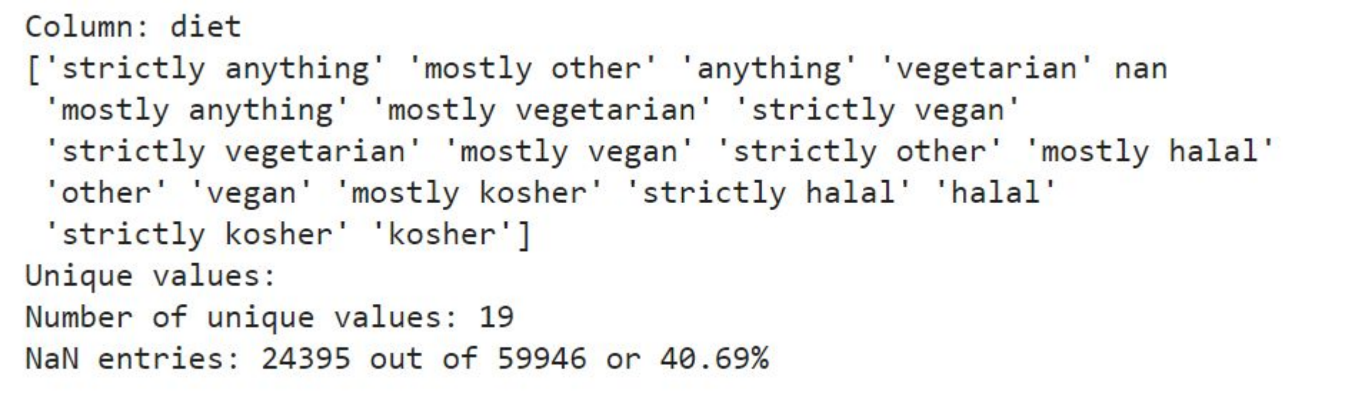
- Goals
  - The goal of this project is to define and attempt to answer one or more questions using machine learning.
  - Given that the data does not contain any information on successful or failed matches to test a model again it does not seem wise to try and predict potential matches or potential categories that could be used to predict matches.
  - Instead I will try to apply a supervised ML model to determine if smoking and drinking predict drug use according to the profile data.
- Data
  - I will use the profiles.csv file supplied from Codecademy and sourced from OKCupid. This file contains the relevant data that appears to be a subset of user data profiles from OKCupid.
- Analysis
  - I will use tools such as numpy, pandas, matplotlib, and seaborne to explore, analyse, and visualise the data. And the LogisticRegression ML model from sklearn to attempt to answer our question(s).
  - Once completed I hope to be able draw our conclusions based the goals for the project.

Data Exploration

- General observations
  - There are 59,946 rows in the file. Each row would appear to correspond to a user profile on the site.
  - There do not appear to be duplicate entries how no unique identifier has been provided.
  - Aside from the pair (orientation/sex) the only columns that gave us any useful information about what someone is looking for in a match are found in the "essay" columns
- Duplicates
  - As no unique identifier has been provided I do not feel that it would be useful to control for duplicates with the aim of removing them. This could be done by select a subset of columns, but we could not be sure that the resulting subset accurately represents the data.
- "Nan" values
  - There are many entries with "NaN" values. The columns range from "diet" to the various "essay0x" columns, and "pets" to name a few.
  - There are in fact only 7 columns out of the 31 that do not contain null entries: ['age', 'income', 'last_online', 'location', 'orientation', 'sex', 'status']

Data Exploration, continued...

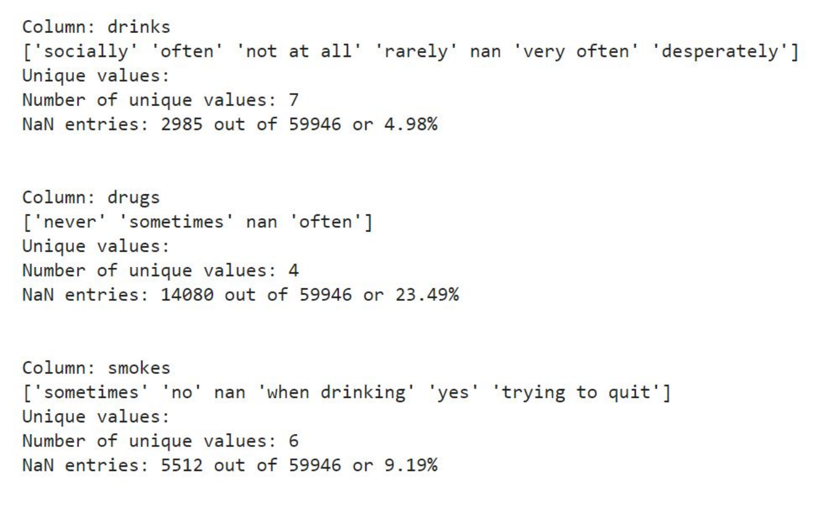
- 'diet' has a relatively high percentage (~40%) of NaN entries and does not seem reliable



```
Column: diet
['strictly anything' 'mostly other' 'anything' 'vegetarian' nan
 'mostly anything' 'mostly vegetarian' 'strictly vegan'
 'strictly vegetarian' 'mostly vegan' 'strictly other' 'mostly halal'
 'other' 'vegan' 'mostly kosher' 'strictly halal' 'halal'
 'strictly kosher' 'kosher']
Unique values:
Number of unique values: 19
NaN entries: 24395 out of 59946 or 40.69%
```

Data Exploration, continued...

- 'drinks', 'drugs', 'smokes', and 'education' each have lower percentage of NaN entries, about ~5%, ~23%, ~9%, and ~11%, respectively

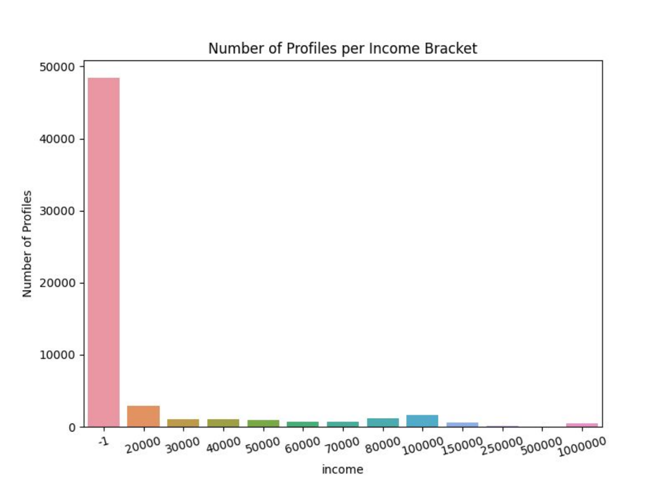


```
Column: drinks
['socially' 'often' 'not at all' 'rarely' nan 'very often' 'desperately']
Unique values:
Number of unique values: 7
NaN entries: 2985 out of 59946 or 4.98%


Column: drugs
['never' 'sometimes' nan 'often']
Unique values:
Number of unique values: 4
NaN entries: 14080 out of 59946 or 23.49%


Column: smokes
['sometimes' 'no' nan 'when drinking' 'yes' 'trying to quit']
Unique values:
Number of unique values: 6
NaN entries: 5512 out of 59946 or 9.19%
```

Data Exploration, continued...

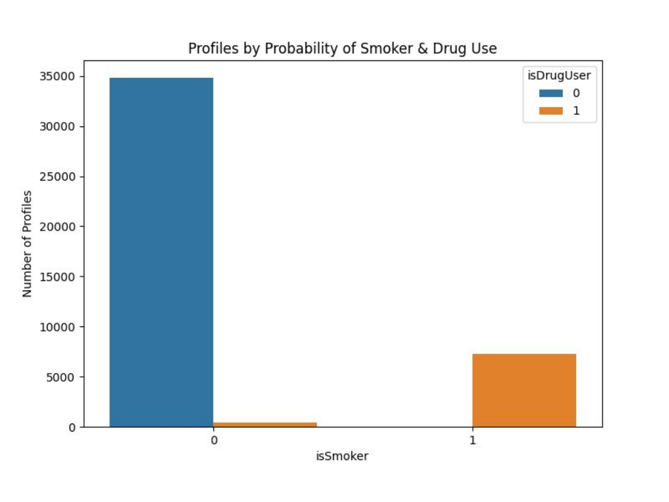
- 'drinks', 'drugs', 'smokes', and 'education' each have lower percentage of NaN entries, about ~5%, ~23%, ~9%, and ~11%, respectively
- 'education' has a relatively high number of possible values (unique) at 33
- 'income' has no NaN entries, but many at (-1) (48,442 entries out of 59,946 total or 80.8%)
  - This is essentially the equivalent of ‘no income provided’ for these rows
  - Given the high rate of (-1) in income this column does not seem reliable

Data Exploration, continued...



- There is an overly large portion of profiles that did not report their income (~48,000)
- I ruled out use of this column (explanation later.)

Data Exploration, continued...



- We can see very clearly that whilst not a lot of profiles that report to be smokers, most of those that do, use drugs.

Data Preparation

- NaN Values
    - Dropped rows whose values in the following columns is NaN: 'drugs', 'smokes', 'drinks'
- Data Augmentation
    - The following columns were added:
        - isDrugUser: value set to 1 where “drinks” corresponds to one of the following values:
            - 'Sometimes' or 'often'
        - isSmoker: value set to 1 where “drinks” corresponds to one of the following values:
            - 'sometimes', 'when drinking', 'yes', or 'trying to quit'
        - isDrinker: value set to 1 where “drinks” corresponds to one of the following values:
            - 'socially', 'often', 'very often', or 'desperately'
    - The code below depicts how these columns were added. Basically a list of potential values that I considered as “drinking” for example were chosen from the list of possible values in the data. For each corresponding value in the list a 1 was assigned in the new column for each matching row.

```
Code:
drinksList = ['socially', 'often', 'very often', 'desperately']
profilesDF['isDrinker'] = profilesDF['drugs'].apply(lambda x: 1 if x in drinksList else 0)
```

Question to answer

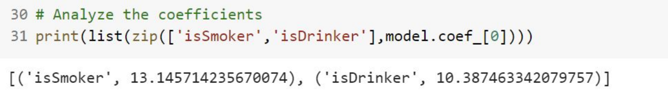
Are users (profiles) who smoke and/or drink more likely to take drugs?

I started by exploring the data with the intent to predict income based on education. However, as I stated in the data exploration section, the majority of profiles did not have any income data. In my mind this made the analysis less interesting.

I then switched to seeing if I could find indicators of “healthy” or “unhealthy’ lifestyle choices. Thus my interest in the columns ‘drinks’, ‘smokes’, and ‘drugs’.

Analysis

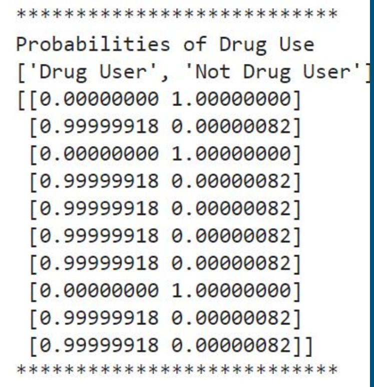
- Logistic regression was used to test if smoking and drinking are indicators of drug use.
- During the training and then testing phase the results (coefficient) show that smoking is the dominant feature that predicts drug use.



```
# Analyze the coefficients
print(list(zip(['isSmoker','isDrinker'],model.coef_[0])))
```

```
[('isSmoker', 13.145714235670074), ('isDrinker', 10.387463342079757)]
```

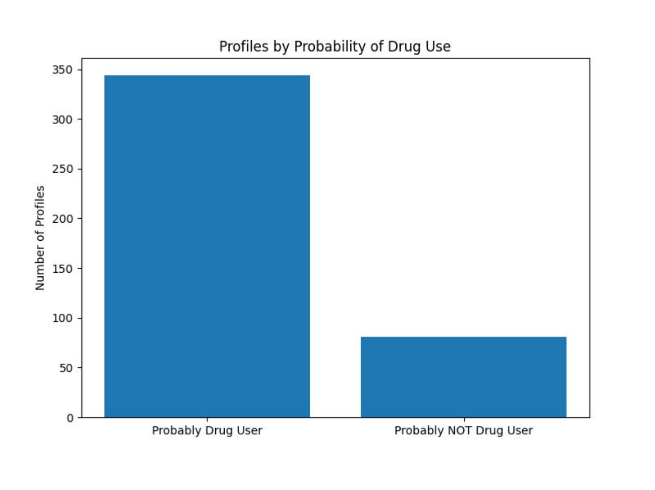
- In the probability predictions I used 80% as the threshold to predict if the profile was likely to be a drug user or not.
- In the prediction subset there were 425 profiles that were split off from the 42,295 rows that remained after dropping the NaN rows and before the train/test phase. 344 had a high probability of being drug users and 81 had a high probability of NOT being drug users (see next slide.)



```
***************************
Probabilities of Drug Use
['Drug User', 'Not Drug User']
[[0.00000000 1.00000000]
 [0.99999918 0.00000082]
 [0.00000000 1.00000000]
 [0.99999918 0.00000082]
 [0.99999918 0.00000082]
 [0.99999918 0.00000082]
 [0.99999918 0.00000082]
 [0.00000000 1.00000000]
 [0.99999918 0.00000082]
 [0.99999918 0.00000082]]
***************************
```

Conclusions

Number of Profiles per Case



- We can see that the majority of the profiles fall in the case of “Probably Drug User”

Are users (profiles) who smoke and/or drink more likely to take drugs?

Answer:

Smoking and drinking are predictors of drug use with smoking being the dominant feature.