Liam Considine
Code Academy Capstone

Cohort: Nov 13, 2018 - Jan 22, 2019

Research Question

My Goal with the OkCupid data set is to determine how well information about the users diet, substance use, age and income can be leveraged to predict each other!

I've tried to make sensible choices, but not do every possible permutation, for instance I might:

Use diet, substance preferences, and age and try to predict income.

or

Use diet, substance preferences and income to try and predict age. Or later try to predict if a user is over 30 years in age!

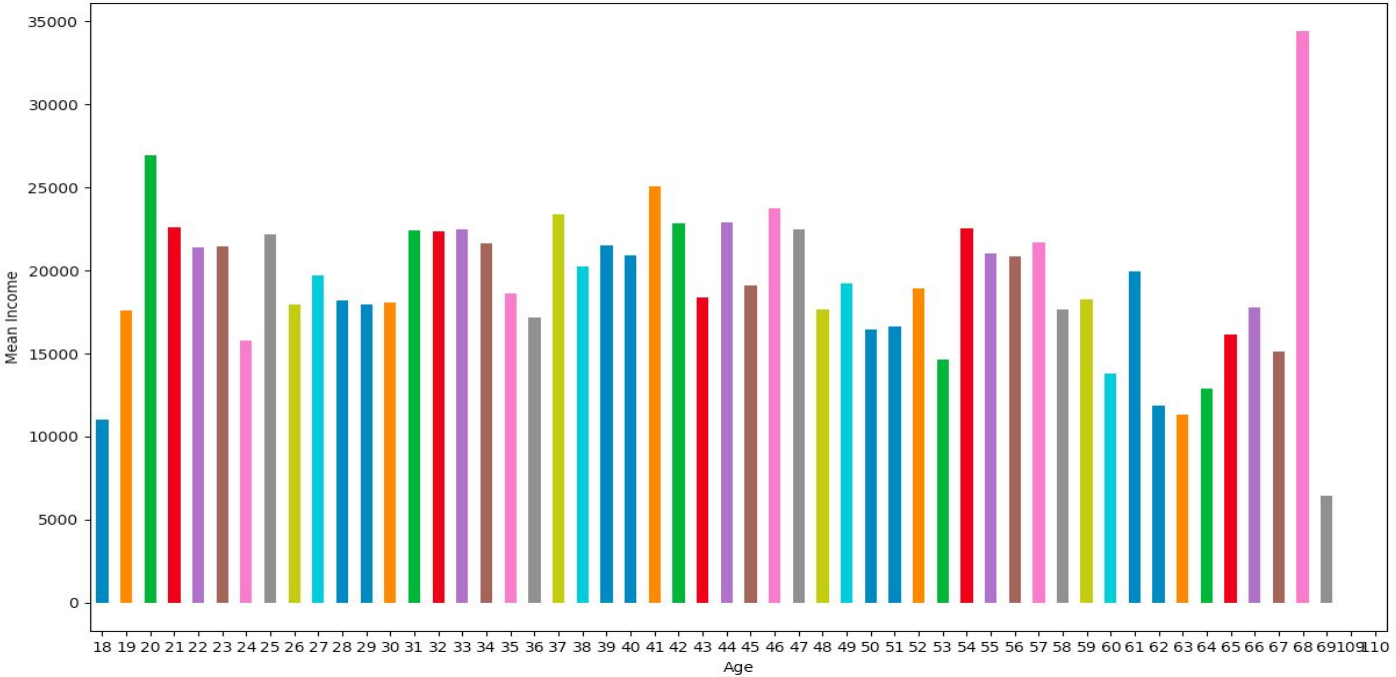
I tried to keep in mind these are users responses about themselves! Selection bias from the start!

Data Exploration

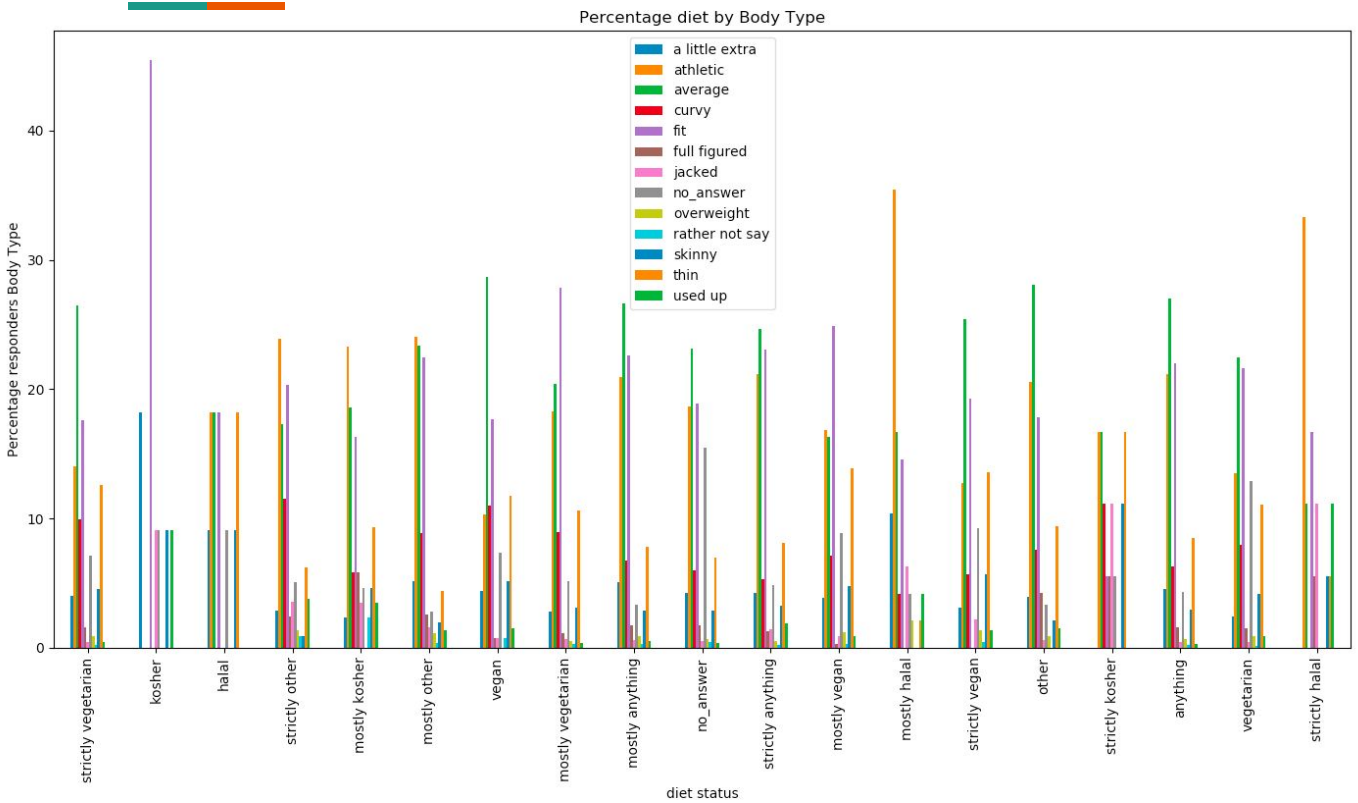
Average income by age: First look seem age isn't going to help much determining income!

The U.S Bureau of the Census has the annual median personal income at **$31,099** in 2016

https://en.wikipedia.org/wiki/Personal_income_in_the_United_States



Data Exploration 2



Users identifying their diet as Halal, tend to call themselves “Athletic”

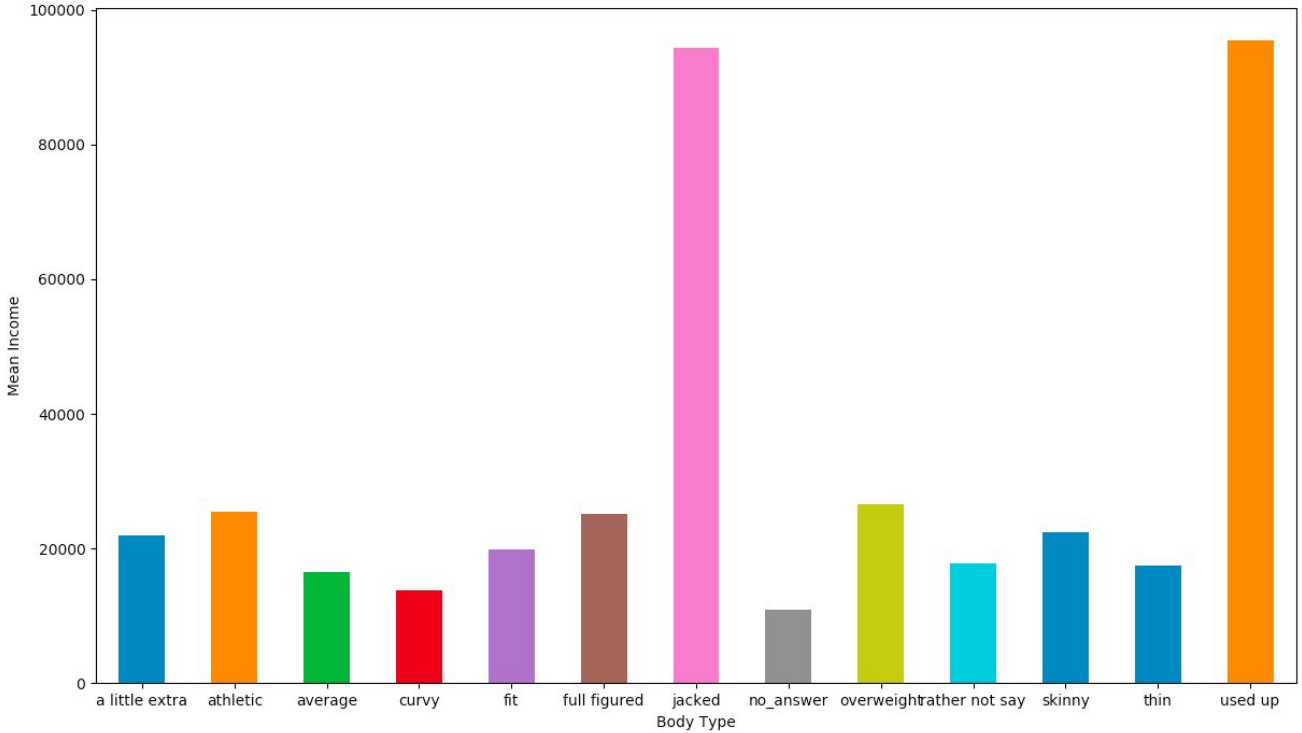
Users identifying their diet as Kosher seem to prefer the similar but semantically different term “Fit”

Vegan, vegetarian and other are mostly “Average”

Data Exploration 3

Users with extreme body types “Used up” for “Jacked” often declare higher incomes on OkCupid.

Jokers? Maybe compensating?



Multiple Linear Regression: continuous value ‘Age’

Independent Variables: "diet", "drinks", "drugs", "body_type", "smokes", “income”

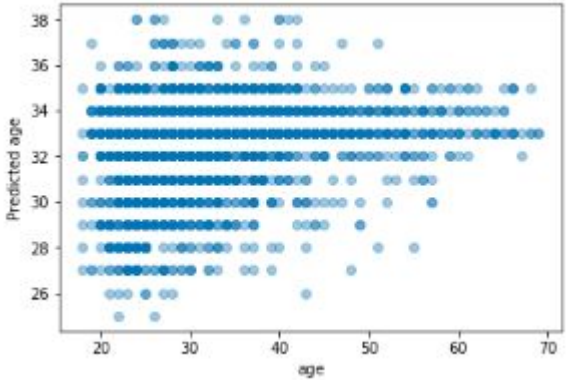
Dependent Variable: “Age”

Coefficients:
[('smokes_code', -5.72685166999203), ('drinks_code', -4.541424576130044), ('drugs_code', -1.7977144151184188), ('diet_code', -1.6844946350432741), ('body_code', 0.3124906365850019), ('income', 0.25781812970865275)]

Model Score:

Train Score: 0.042332739364382754

Test Score: 0.04995830280401504

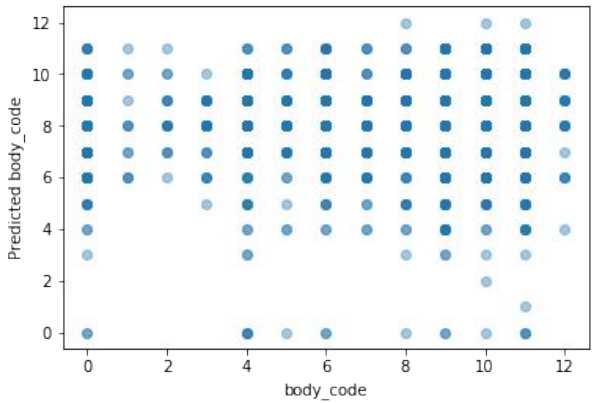


Made sure to round the models predictions before checking the models score. Still does really poorly!

The model never predicts over 38 years of age!

What if I checked if the model predicted within 5 years of the users actual age? I’d probably have to write my own model.score function. Could be a useful exercise!

KNeighborsRegressor: continuous value 'body_code'




```
body_type_map = {
    "no_answer": 0,
    "rather not say": 1,
    "used up": 2,
    "overweight": 3,
    "curvy": 4,
    "full figured": 5,
    "a little extra": 6,
    "skinny": 7,
    "thin": 8,
    "average": 9,
    "fit": 10,
    "athletic": 11,
    "jacked": 12
}
```

I mapped peoples' body_type to a new feature body_code. Focused on keeping the scale 0 "least fit" body type to 12 "most fit". I then asked a KNeighborsRegressor to predict the body_code… and rounded to the nearest integer. Treating body_type as a continuous value did not prove to be a good option!

I think including 'no_answer' and 'rather not say' could also be misguided.

The model predicts a lot of 'Average' people! In the data i correctly labels ~29% of all average (body_code 9) instances *recall*, but when it predicts (body_code 9) it is only correct ~26% of the time *precision*.

Input data shape: (10998, 6)
Independent Variables: income, age, diet_code, drinks_code, smokes_code, drugs_code
Dependent Variable: body_code
Score: -0.13684386635145618

	precision	recall	f1-score	support
0	0.12	0.02	0.03	112
1	0.00	0.00	0.00	12
2	0.00	0.00	0.00	17
3	0.00	0.00	0.00	33
4	0.07	0.02	0.03	118
5	0.02	0.02	0.02	47
6	0.07	0.05	0.06	159
7	0.05	0.17	0.08	75
8	0.07	0.31	0.11	135
9	0.26	0.29	0.28	589
10	0.23	0.18	0.20	419
11	0.21	0.05	0.08	459
12	0.00	0.00	0.00	25
avg / total	0.18	0.15	0.15	2200

KNeighborsClassifier: binary feature categorization

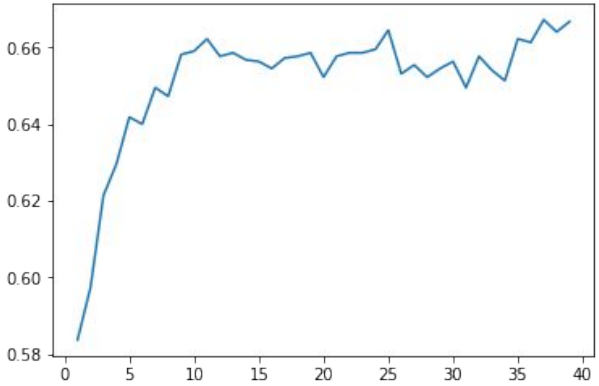
I fed a KNeighborsClassifier the questions below. In both cases I used the same dependent variables income, diet, drugs, drinks, smokes, age (minus the parameter we are choosing to be independent!) Checked the model score with n_neighbors between 1 and 40.

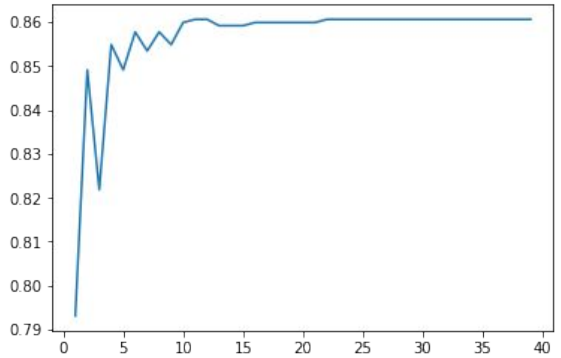
1) Can you predict if a user is over 30 years of age

The model does fairly well, but still only around 68%

2) Can you predict if a user is a mostly vegetable eater?

The model does well! But alas, the data is highly unbalanced… almost 6 out of 7 users eat anything or are meat heavy! Doing some more work on my dataset might improve this result!



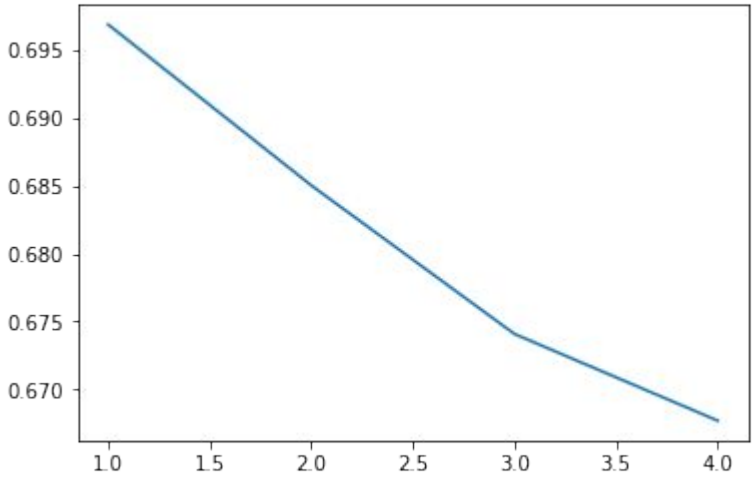


SVC: binary feature categorization

I ran the same question about users being over 30 years of age by a SVC (support vector machine) model. With gamma between 0.5 and 4 for the radial bias function.

Interestingly a low gamma scores best, and quickly drops off, this means making the model less sensitive to the training data creates the best results.

When the model predicts a user is over 30 years old, it labeled 72% of those cases correctly: *precision*. However, it only labeled a user over 30 69% of the cases where a user is over 30 that exist in the dataset *recall*.



	precision	recall	f1-score	support
0	0.67	0.71	0.69	1038
1	0.72	0.69	0.71	1162
avg / total	0.70	0.70	0.70	2200

Summary

All in all a great exercise. Things I learned.

1) Do not jump right in and start modeling! Create a question and the scope of a question that you know you can undertake with a variety of approaches and focus on. I ended up all over the place!
2) Using linear regression mapped onto “continuous categories” is not a really good idea. I found this dataset difficult to use for these cases because even “income” was ranges ($0 - $20,000)
3) Really take the time to look through all the data up front, I jumped right in and said these are the columns i’m gonna use cause I need to get through this assignment!
4) Focus on the question to lead to the right algorithms to experiment.

Please give me as much feedback as possible!

-liam