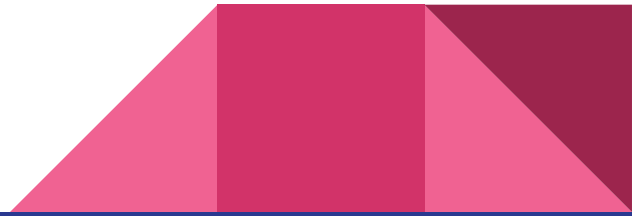
Body-image & Income in Online Dating on OKCupid

Machine Learning Fundamentals

Layla Messner

11/05/2018

Table of Contents



My Question

Does body-image predict income?

I thought body-image (the way a person sees and describes their body) and income might both be markers of self-esteem, so I wanted to explore what relationship, if any, might exist between body-image and income.

Biases & Limitations

Gender Binary

All individuals in the OKCupid dataset are classified as male or female.

As such, the data is missing information about people who experience gender in a non-binary way.

Skewed Dataset

There were far more men than women in the dataset.

- Since the dataset for women was so small, I focused on men for the most part, introducing more bias.

Word Connotation

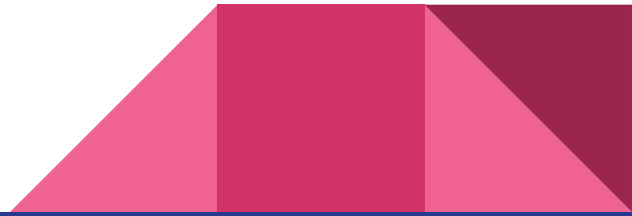
When creating the data about body-image, I classified based on the connotation of the body-type descriptor. While these decisions were based on 10 years of experience as a fiction author, this was not a scientific process.

Hmm...

There are significantly more men than women in the OKCupid dataset. I wonder if OKCupid is marketted towards men, or if men are more likely to online date, in general?

(While these questions are interesting to ponder, they are beyond the scope of this project.)

To answer my question, I used pandas, numpy, matplotlib, and scikit-learn. I began with augmenting the data.



The OKCupid dataset included the following info:

- Age
- Body-type
- Diet
- Drinks
- Drugs
- Education
- The text from 9 essays
- Ethnicity
- Height
- Income
- Job
- Last-online
- Location
- Offsprint
- Orientation
- Pets
- Religion
- Sex
- Sign
- Smokes
- Speaks
- Status

Data Augmentation: New columns using .map()

Body-Image → Income Brackets → Income Binary

Body-image classes:

2 (positive): curvy, fit, thin, athletic, full-figured, jacked

1 (neutral): average, a little extra, skinny

0 (negative): used up, overweight, rather not say

Income-brackets classes:

0 (under median): 29K or less

1 (5-figures): 30K - 99K

2 (6-figures): 100K - 999K

3 (7-figures +): 1 million and up

Income-binary classes:

I also divided the dataset into two income classes:

0: Those making less than the median individual income for 2012 ($28,213)

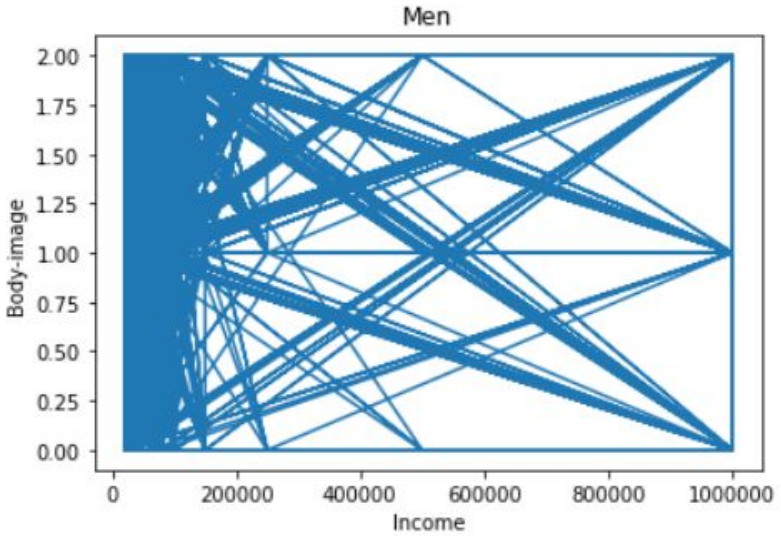
1: Those making more

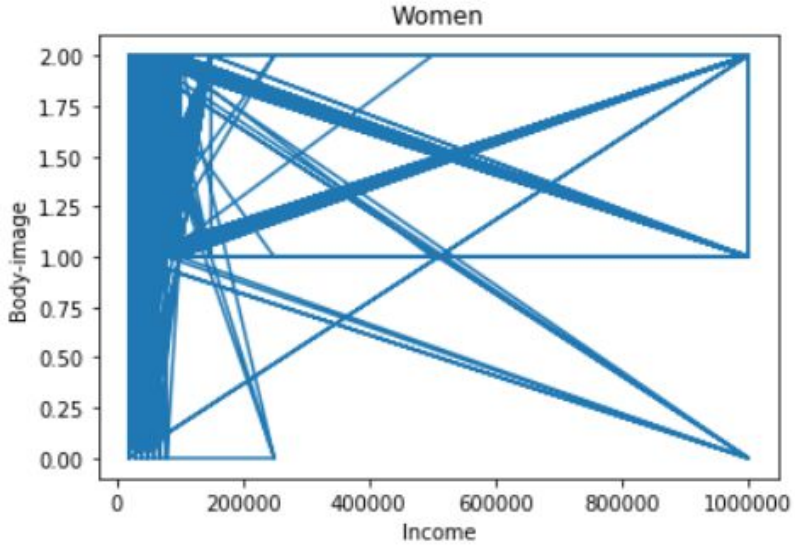
Here’s a sample from my cleaned and augmented dataset:

	body-type	income	essay-len	body-image	Income-bracket	Income-binary
Man 1	average	80,000	1453	1	1	1
Man 2	thin	20,000	477	2	0	0
Man 3	average	40,000	4629	1	1	1
Man 4	fit	60,000	6938	2	1	1
Man 5	athletic	20,000	397	2	0	0

I normalized this data using scikit-learn’s min-max-scaler. (Later, when performing classification, I had to revert the income data to discrete classes.)

Exploring the Data: Visualizing income vs body-image with pyplot line graphs





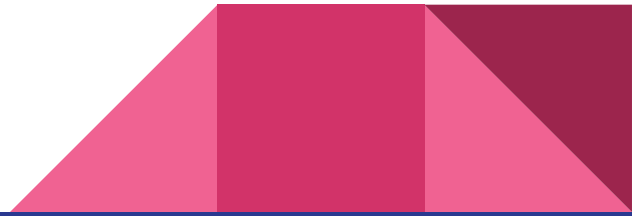
It looks like women may be less likely to use words that indicate negative body image. However, this could also be the result of having less female data in the higher income ranges.

Machine Learning with scikit-learn

To answer my question, I used the following approaches:

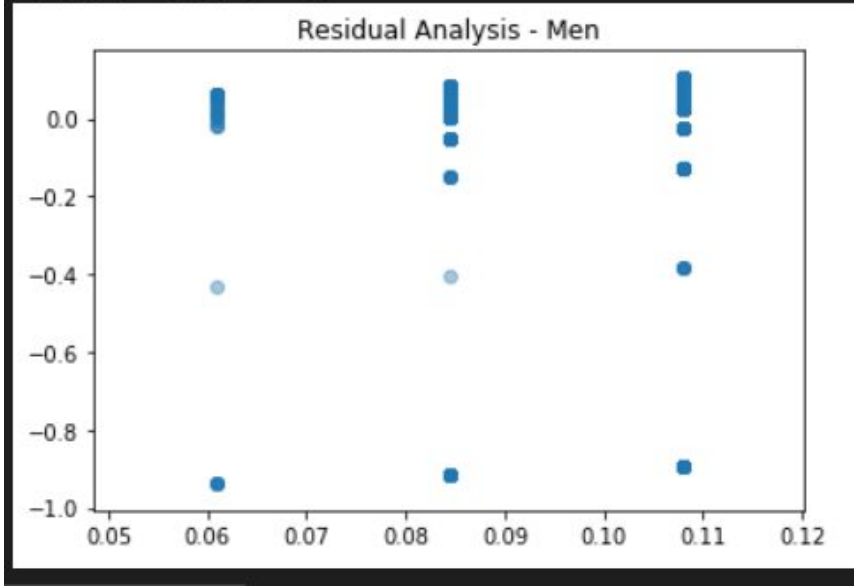
1. Regression
    a. Single Variable Linear Regression
    b. Multiple Linear Regression
    c. K Nearest Neighbors Regressor
2. Classification
    a. K Nearest Neighbors Classifier
    b. Sports Vector Machine

The time required to run each of these models was negligible.

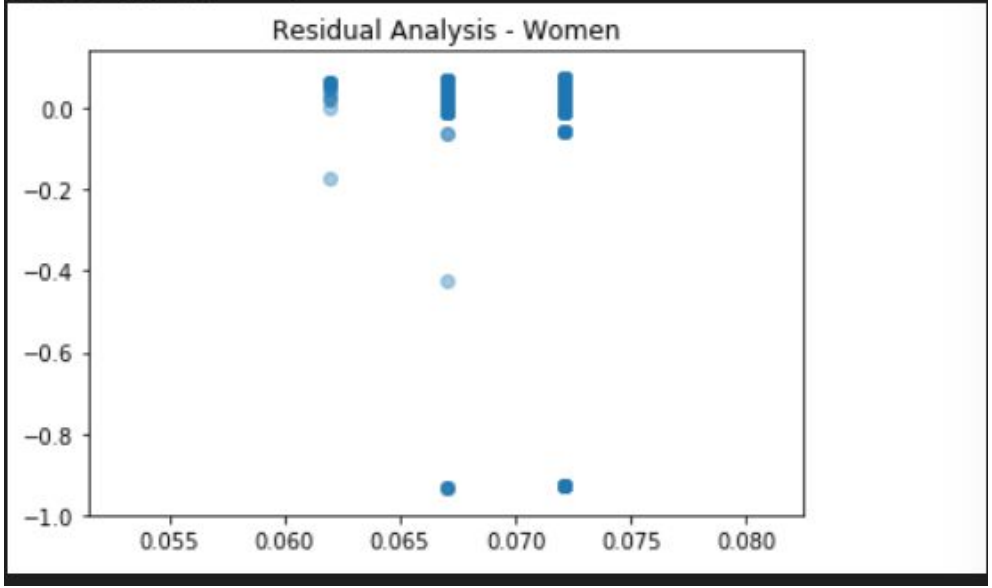


Regression

```
Train score:
0.0036830678502619563
Test score:
-0.000491158567641703
```

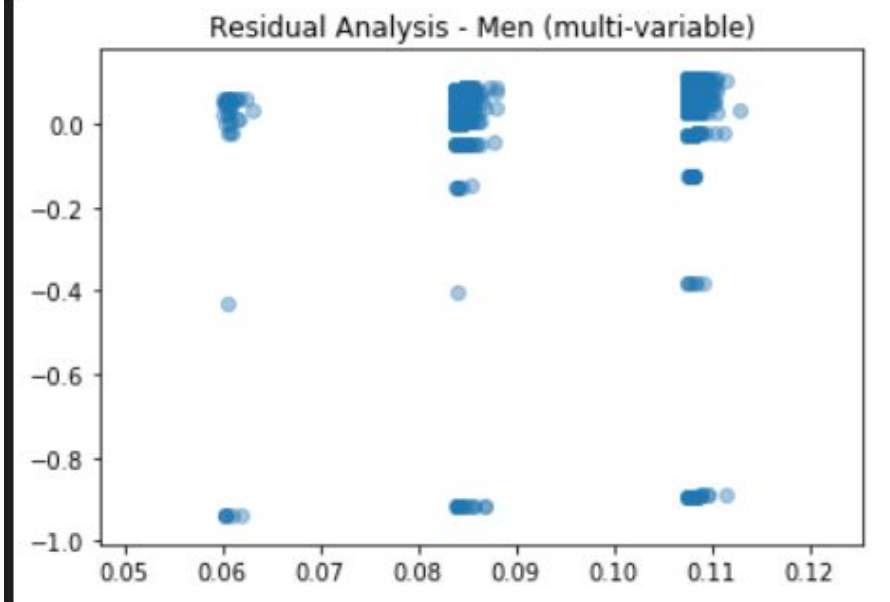



```
Train score:
0.00018725258103879927
Test score:
-0.003265619037417622
```

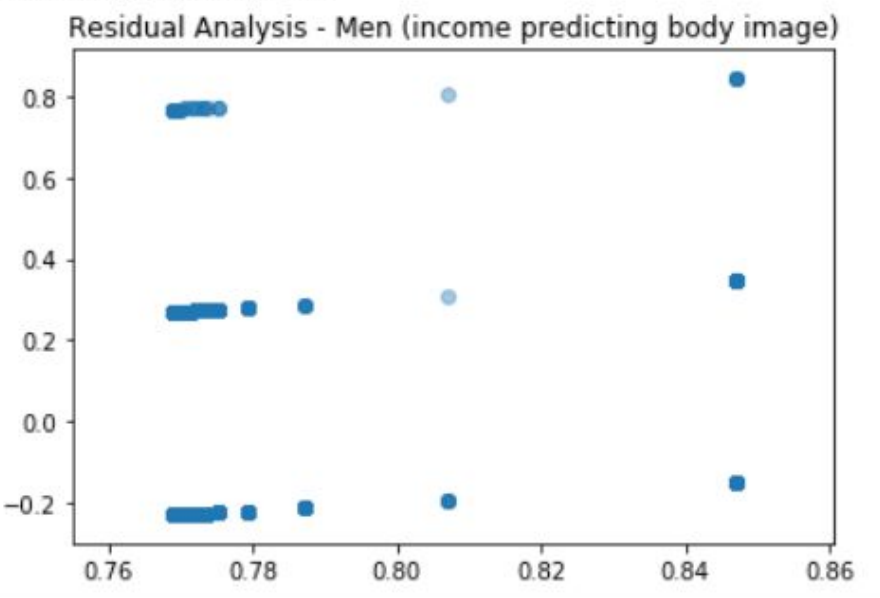


Using scikit-learn, I tried both Single and Multi variable linear regression, for the data on men and women (separately), looking at body-image as predictor of income, and visa versa.

```
Train score:
0.003692341399568755
Test score:
-0.0003095986165702058
```




```
Train score:
0.0036830678502620673
Test score:
0.0004467530347510573
```

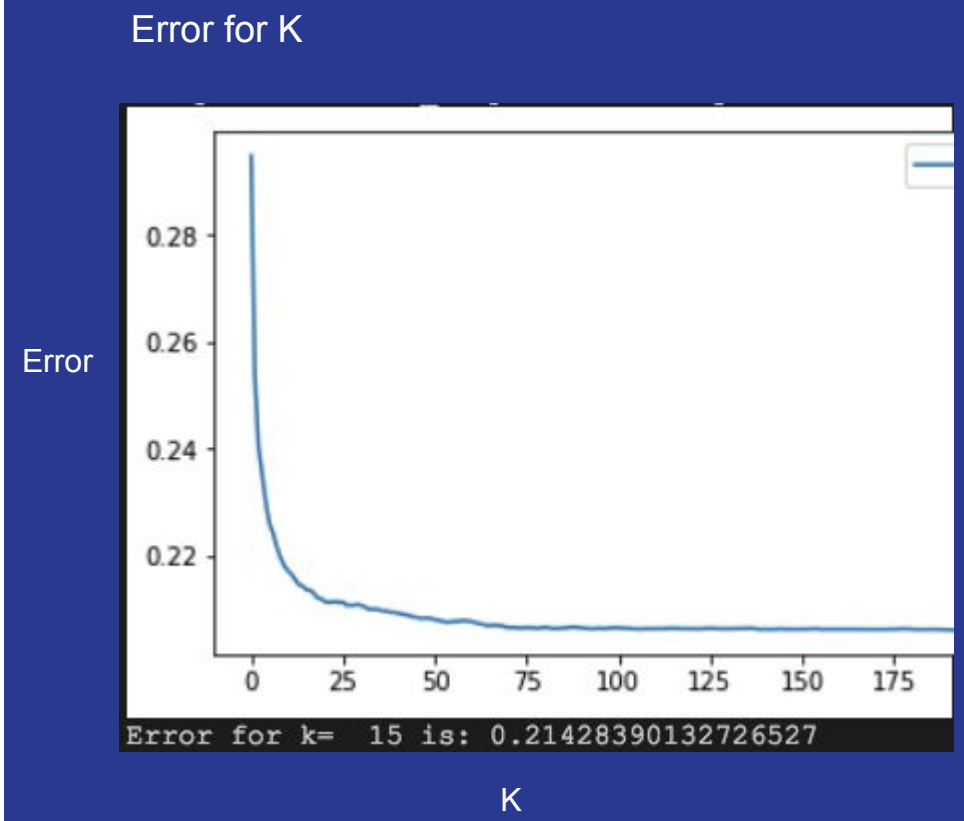


In all cases, the train scores were low, test scores were even lower - and wildly different.

None of these models were able to accurately predict y.

KNN Regressor

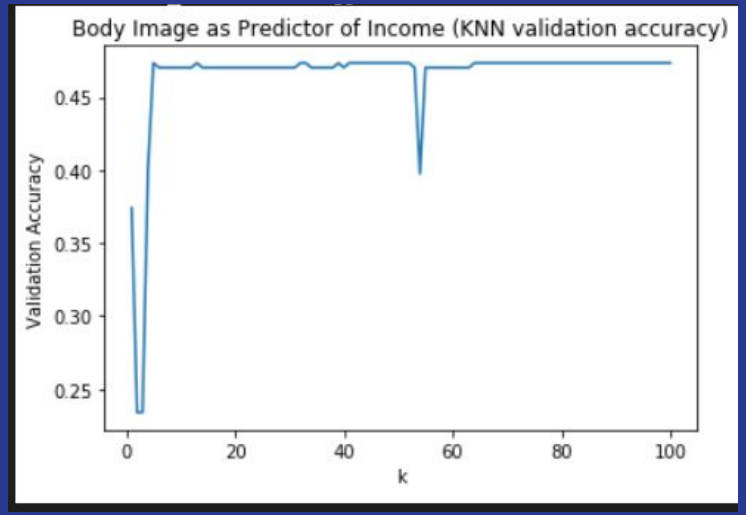
Scikit-learn's KNN Regressor algorithm also had unimpressive accuracy predicting income from body-image, though it was more accurate - and simpler to run - than Linear Regression.



Classification

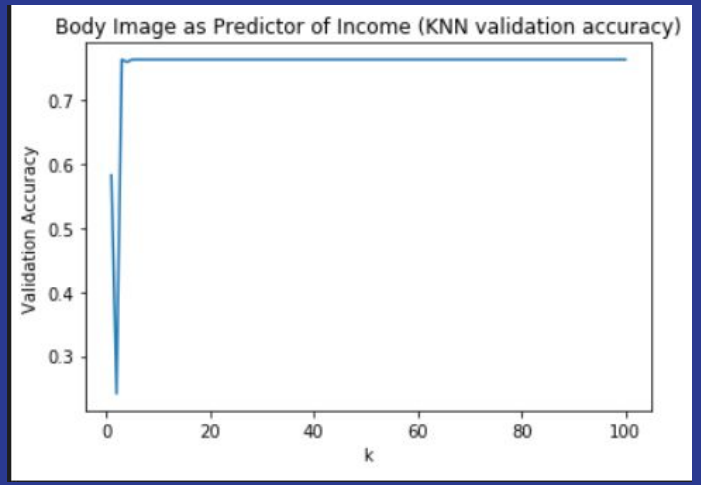
KNN - The most accurate model

I was excited when I switched to KNN classification using only 2 income classes, and accuracy increased to 76%...



KNN Accuracy with 4 Income Brackets

KNN Accuracy with only 2 Income Classes



...but then I calculated Precision, Recall, & F1 score

KNN Classifier:

Accuracy: 77%

Precision: 38%

Recall: 50%

F1 Score: 43%

It turned out the model had reached 76+% accuracy by predicting “1” as the value for income-binary, for every individual in the validation set.

In other words, the model predicted that everybody would have an income over the individual median income - and was accurate 76-77% of the time.

Obviously, this is not a valid model.

What about Sports Vector Machine?

I ran scikit's SVC algorithm with *exactly* the same results as KNN, in terms of accuracy, precision, and recall. Clearly, the SVC model also predicted a "1" as income for every individual in the dataset.

SVC and KNN were comparable in terms of simplicity and time to run, as well.

At this point, I felt I had enough information to answer my question.

In conclusion...

There is no relationship
(dating pun not intended)

To my surprise, I see no evidence of a relationship between body-image and income.

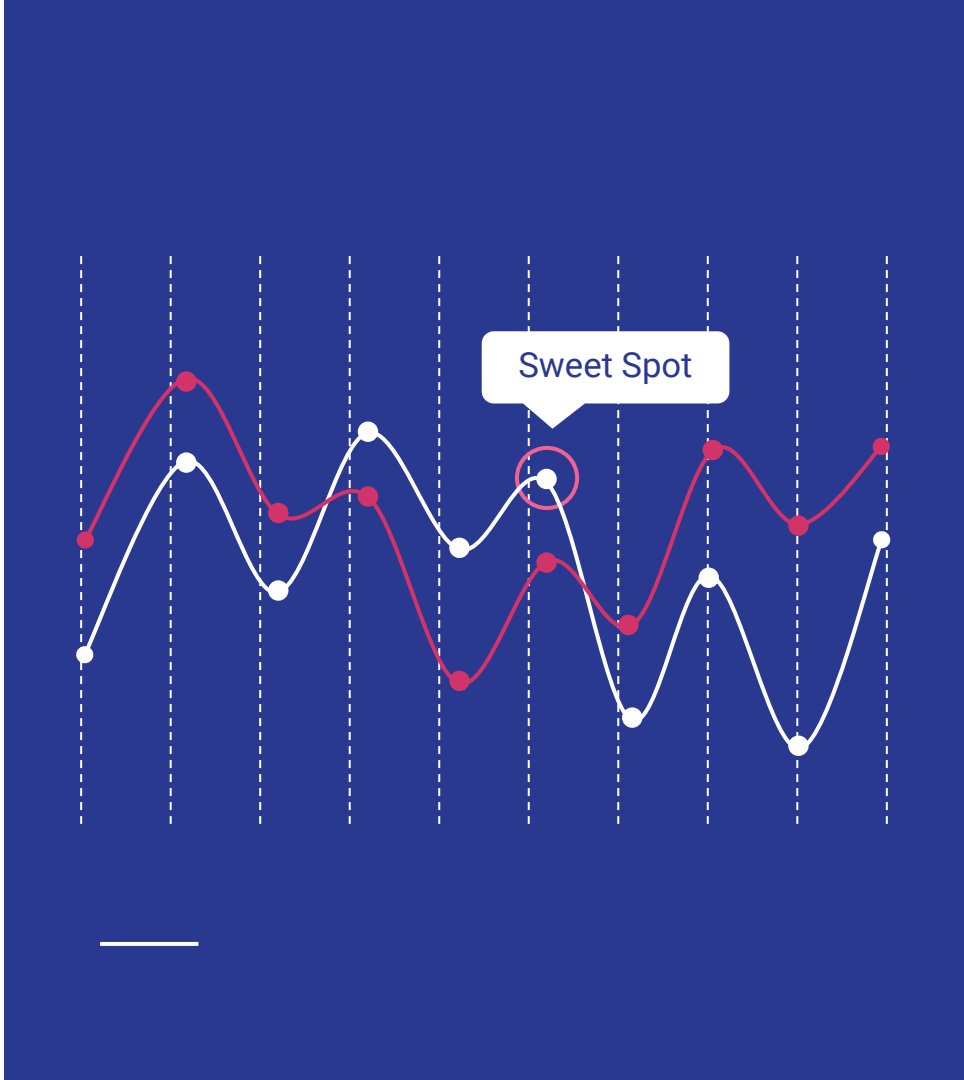
This may mean that one (or both) of those factors does not accurately reflect self-esteem, that self-esteem has no bearing on income, or that physical-esteem and professional-esteem are distinct.

Good News:

A man’s earning potential is not affected by his body-image.

Further Data Needed

Which members of this dataset found partners on OKCupid? How long did that take them? How long did the relationships last?



Next Steps

1. Explore the current data for body-image and income using Unsupervised Learning.
2. Obtain further data about the outcome of each person's time on OKCupid - did they find a partner? How long did it take? How long did the relationship last?
3. Explore any links between body-image and/or income and relationship-success on OKCupid.
4. Find out if more men than women online date (to satisfy my own curiosity).

Read the code at:

https://github.com/LaylaUX/online-dating