Date A Scientist

Machine learning with the OKCupid data set

Andrew Hercules | 22 Jan 2019

Introduction

Preliminary research

There is a wide and diverse range of research on the links between body type, diet, lifestyle factors, and age.

Some research findings are well-known (e.g. smoking suppresses appetite and is decreasing amongst young people)

Other findings aren't so well known (e.g. drug use increasing in baby boomer generation and not just confined to those deemed "unhealthy")

- Body changes as we age
- Changing attitudes towards drugs within baby boomer demographic
- Smoking rates are falling among young people
- Smoking suppresses appetite and "can make you skinny"
- Alcohol consumption can cause weight gain

Research questions

Based on my preliminary research, I developed two research questions as I wanted to know if I could build a model that would utilise lifestyle information and diet to predict age and body type.

1. **Can we use diet, lifestyle information (alcohol consumption, drug use, tobacco use), and body type to predict age?**

2. **Can we use diet, lifestyle information (alcohol consumption, drug use, tobacco use), and age to predict body type?**

Rationale for research questions

Predicting age could be interesting for various reasons, including:

- **Advertising:** marketing companies are known to covet the 18 - 35 year old demographic and predicting age could be useful if providing age is not a mandatory requirement during the sign-up process
- **False profiles:** predicted age - along with facial recognition - could be useful in determining if someone younger than 18 is attempting to use the service (e.g. if their answers make them seem like they are 40+ but facial recognition detects they are a teenager)

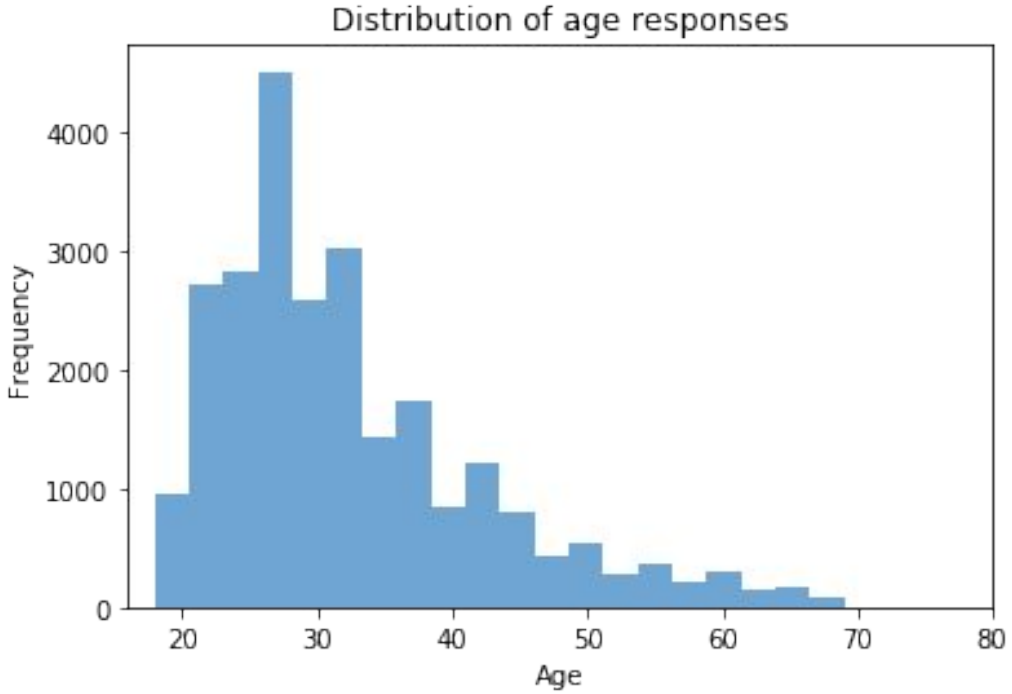
Predicting body type could also be used for marketing purposes as companies may want to target specific users based on body type (e.g. fitness apparel companies targeting those with an “athletic” body type)

Exploring the data: Distribution of ages

Key trends:

Most users between the coveted 18 to 35 demographic

No members younger than 18 and older than 70

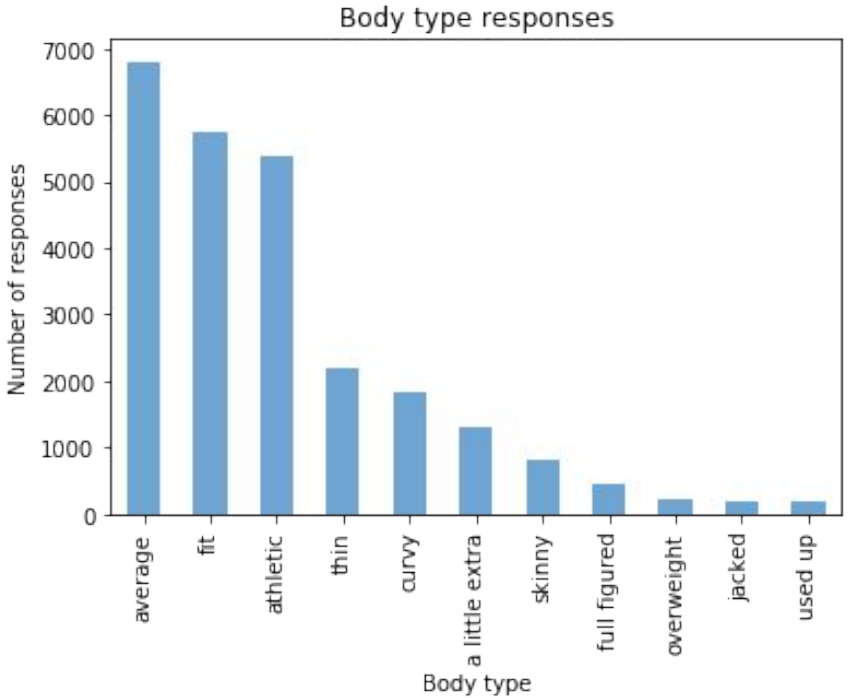


Exploring the data: Body type responses

Key trends:

Most popular response is “average” while least popular response is “used up

Considerable number of “average”, “fit”, and “athletic” responses could skew results



Preparing and cleaning the data

Removed NaN (not a number) values, which resulted in **a drop of more than 50% of responses** - from 59,946 to 25,124

Created new columns to code categorical data

- Diet (diet → diet_code)
- Body type (body_type → body_type_code)
- Tobacco use (smokes → smokes_code)
- Alcohol consumption (drinks → drinks_code)
- Drug use (drugs → drugs_code)

Removed columns that would not be used for analysis, including columns with essays, income, job, education, and location

Question 1:

Can we use diet, lifestyle information (alcohol consumption, drug use, tobacco use), and body type to predict age?

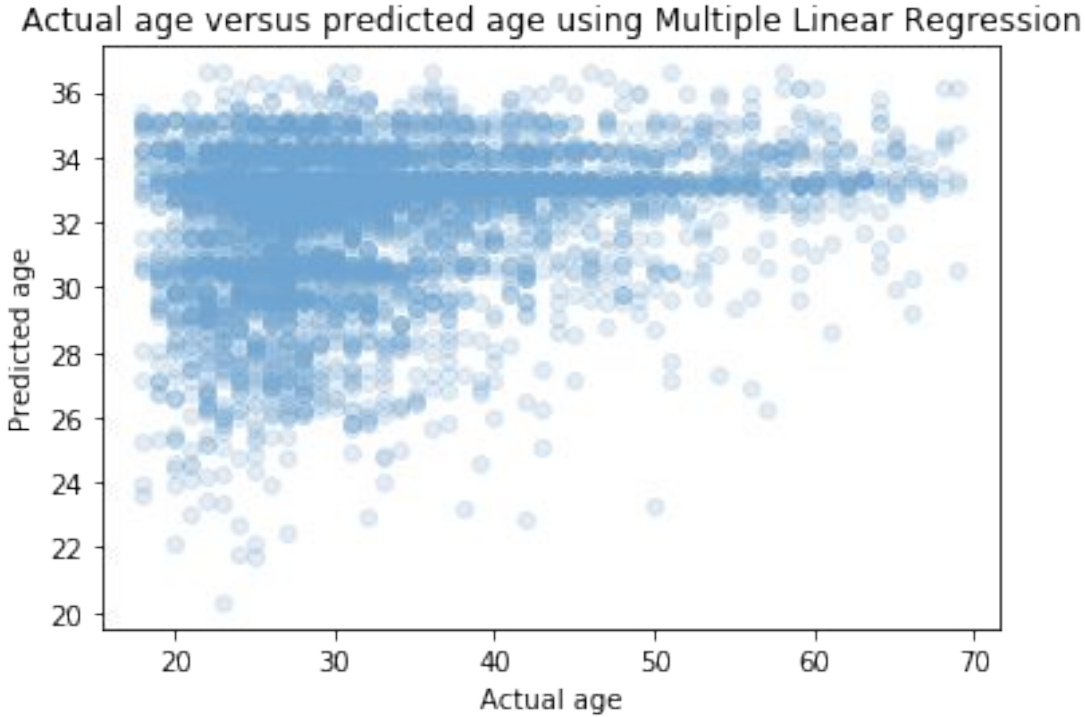
Predicting age - Multiple Linear Regression

Training set score:

0.0463680266632

Test set score:

0.0432589272067



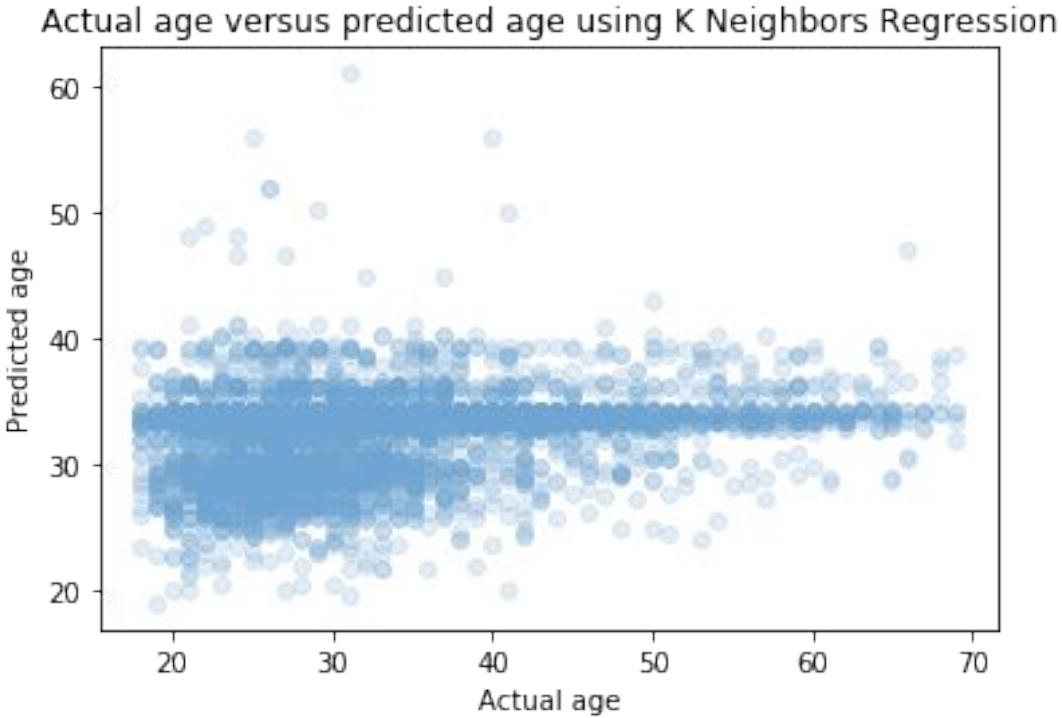
Predicting age - K Neighbors Regression

Training set score:

0.09228919445

Test set score:

0.026069192568



Predicting age - method comparison

While the K Neighbors Regression was more accurate with a higher R^2 score for the training set, the Multiple Linear Regression produced a higher R^2 score for the test set - but both methods did not produce a model with a high accuracy

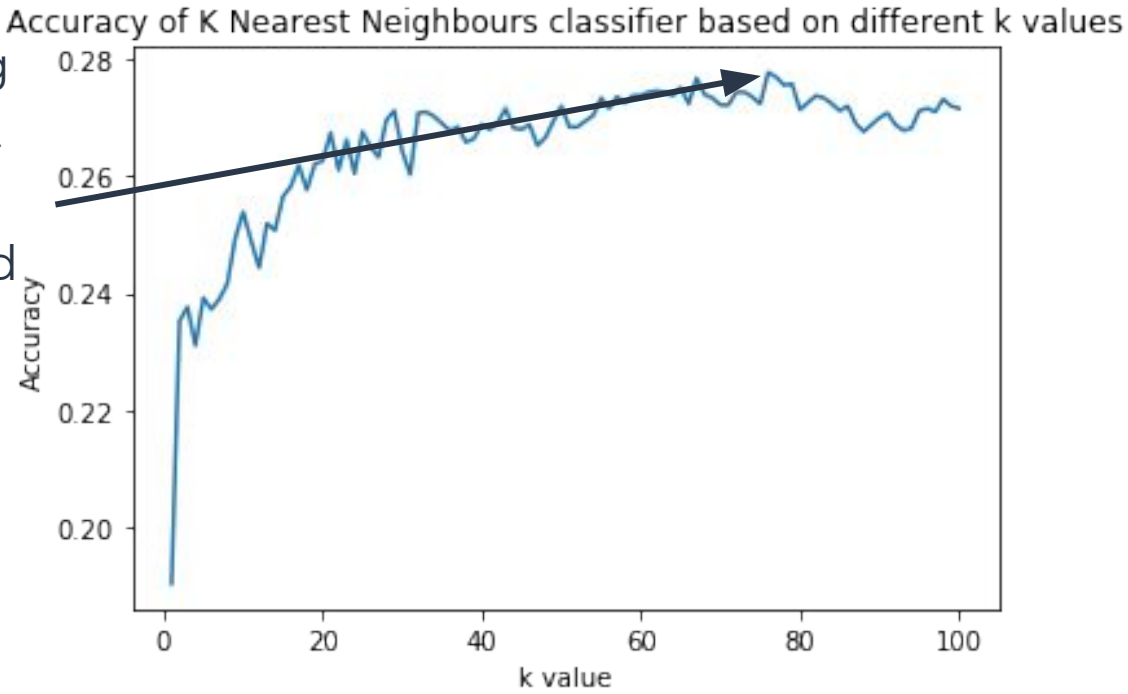
Metric	Multiple Linear Regression	K Neighbours Regression
R^2 score - training	0.0463680266632	**0.09228919445**
R^2 score - test	**0.0432589272067**	0.026069192568

Question 2:

Can we use diet, lifestyle information (alcohol consumption, drug use, tobacco use), and age to predict body type?

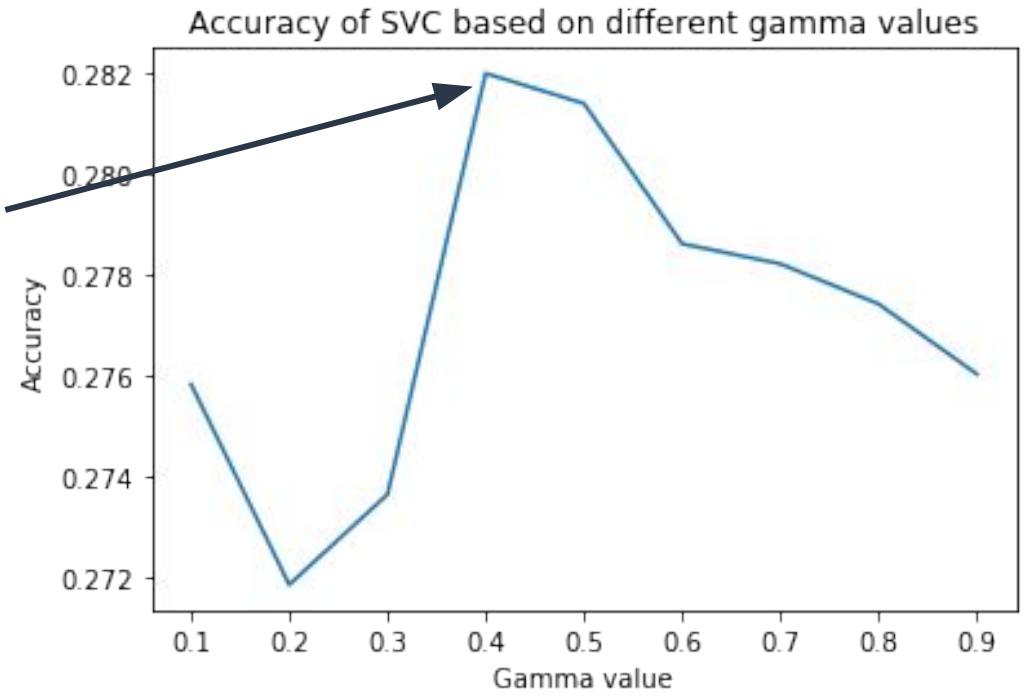
Predicting body type - K Nearest Neighbors

With training data, highest accuracy of 0.27781 found when k = 76



Predicting body type - Support Vector Machines

With training data, highest accuracy of 0.28199 found when gamma = 0.4



Predicting body type - method comparison

The Support Vector Machine classifier had higher accuracy and precision values whereas the K Nearest Neighbor classifier had higher recall and F1 score values

Metric	K Nearest Neighbor	Support Vector Machines
Accuracy	0.27781	**0.28199**
Precision (overall)	0.19	**0.24**
Recall (overall)	**0.28**	0.26
F1 Score (overall)	**0.22**	0.21

Conclusions

Conclusions

When predicting age, there was a very low coefficient of determination score when using both Multiple Linear Regression and K Means Regression, indicating a weak association between the features selected for analysis (diet, lifestyle information, body type) and age.

When predicting body type, the Support Vector Machine method had an accuracy score of 28%, slightly higher than the K Nearest Neighbor score of 27.7%.

Conclusions - a personal note

You may wonder why I chose to continue building machine learning models that resulted in low accuracy scores.

I think it's very important to not only show the times when machine learning is successful in predicting something, but also when it's not successful. You can learn just as much by studying why something worked and why something did not.

In this case, I suspect that predicting age can be challenging as people of all ages come in all different shapes and sizes with different diets and lifestyle choices - and so it's difficult to find patterns. Perhaps a more powerful unsupervised machine learning method would fare better ...

Next steps

Consider re-running analysis with an unsupervised algorithm to see if it can uncover patterns to predict age and body type

Rather than predicting the exact age, consider grouping into bands (e.g. 20-30, 30-40, etc.) to predict the potential age range

Try and obtain more complete data - while processing and cleaning the data, I had to drop more than 50% of the rows due to no responses

Consider that the features selected (body type, age, lifestyle information, diet) are all self-reported and so there is possibility that the responses will suffer from self-reporting bias

Thank you! :-)