Date-A-Scientist

Machine Learning Fundamentals

Jacopo Martolini

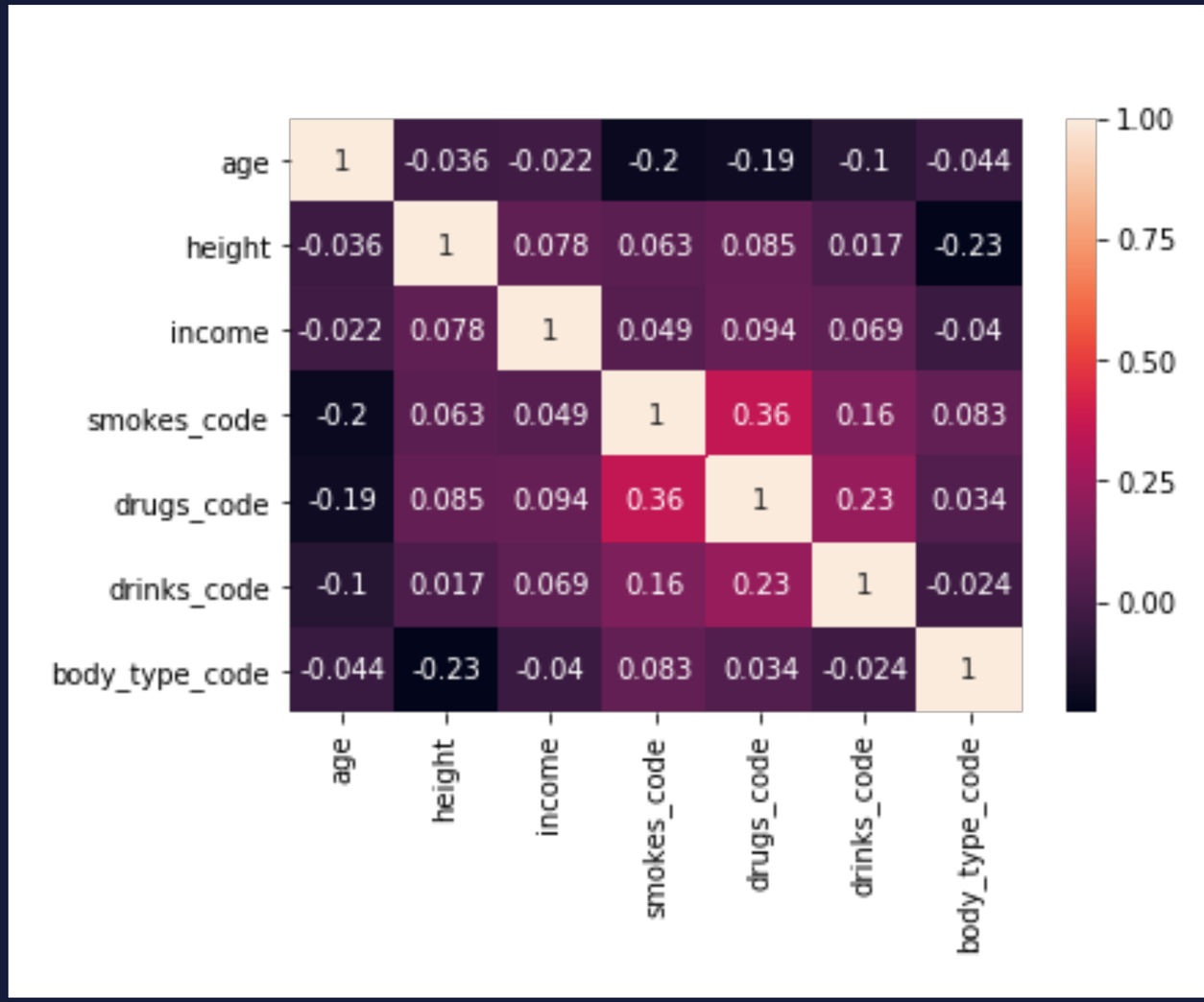
Table of Contents

- Exploration of the Dataset
- Questions to Answer
- Augmenting the Dataset
- Classification Approaches
- Regression Approaches
- Conclusions/Next steps

Exploration of the Dataset

The dataset is composed of various features of OkCupid users profiles. This self assessment data should have a data cleansing beforehand.

It looks like there isn’t much correlation between features.

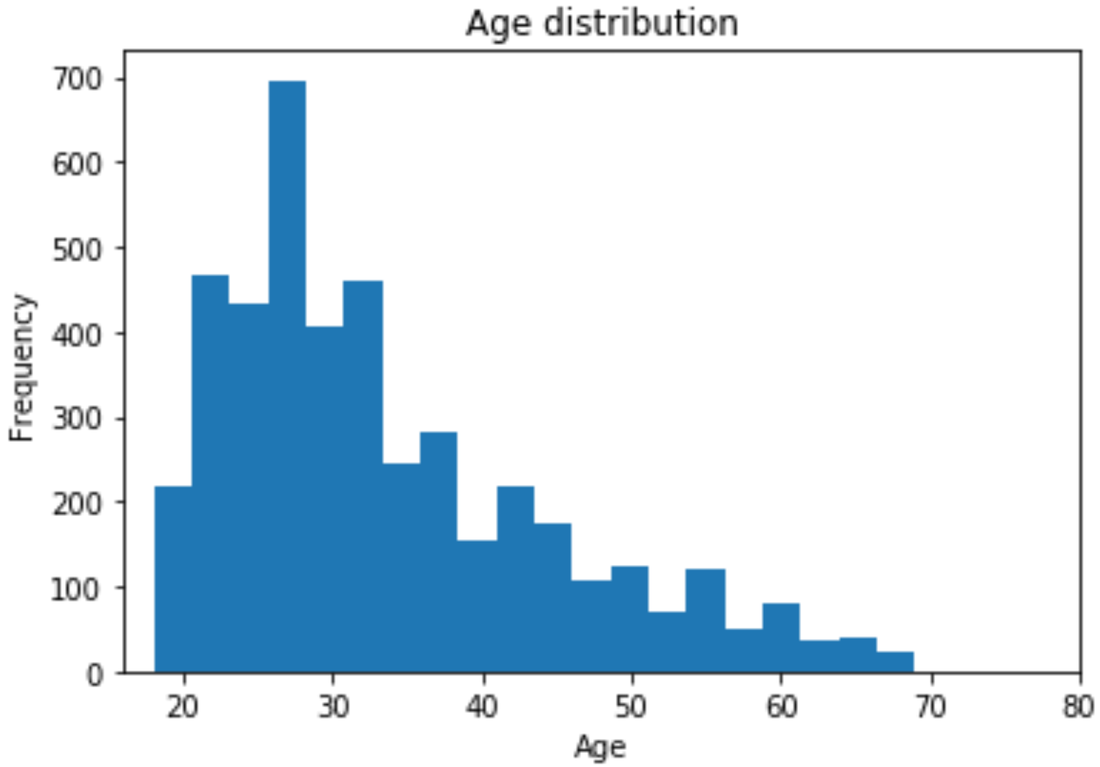


Exploration of the Dataset

The user base age is right skewed.

The mean is 33 years old with a standard deviation of 11.

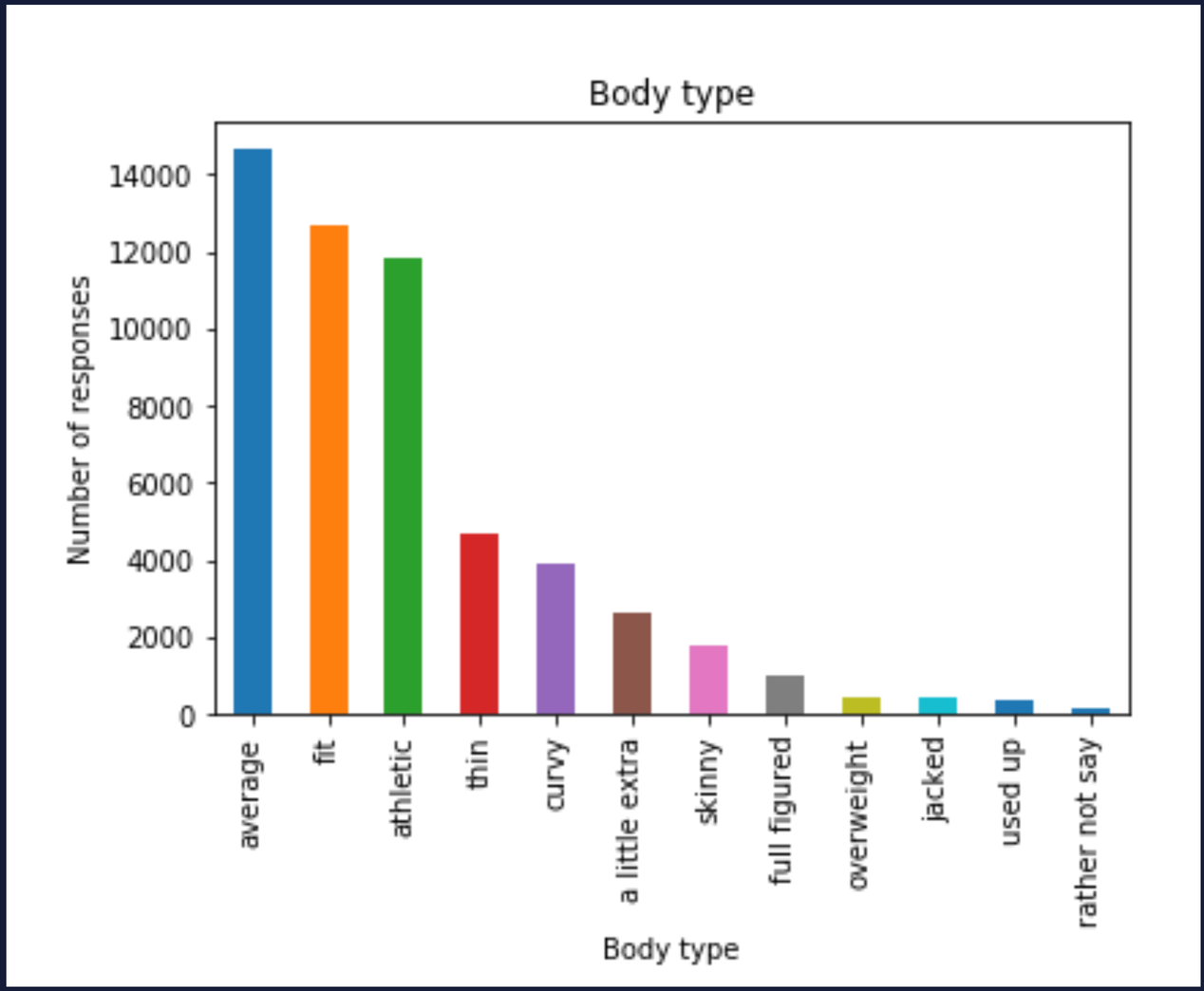
The lower limit is 18, imposed by terms and conditions of this social network.



Exploration of the Dataset

Being self assessed, the body type feature is subject to lack of accuracy.

However we can observe a general tendency to a fit user base.



Questions to Answer

I would like to explore the correlation between smoke, drugs abuse and body type.

From the analysis of the data set it looks like there is a feeble negative correlation between age and recreational substances. I would like to understand if age can be predicted by smoke, drinking and drugs use.

Augmenting the Dataset

In order to understand the correlations, these values needed to be mapped:

- Smokes
- Drinks
- Drugs
- Body type

I've dropped values that were not very number relevant, another way could have been mapping them to similar values.

Once mapped the columns are added to the dataframe.

```
smokes_mapping = {
"no": 0,
"trying to quit": 1,
"when drinking": 2,
"sometimes": 3,
"yes": 4
}

df["smokes_code"] = df.smokes.map(smokes_mapping)

to_drop = ['skinny','full
figured','overweight','jacked','used up','rather not say']
df = df[~df['body_type'].isin(to_drop)]


df["body_type_code"] = df.body_type.map(body_type_mapping)
```

Classification Approaches

Predicting body type from smoke and drugs abuse using Naive Bayes and K-nearest Neighbors Classifier

The first model has a score of 0.305

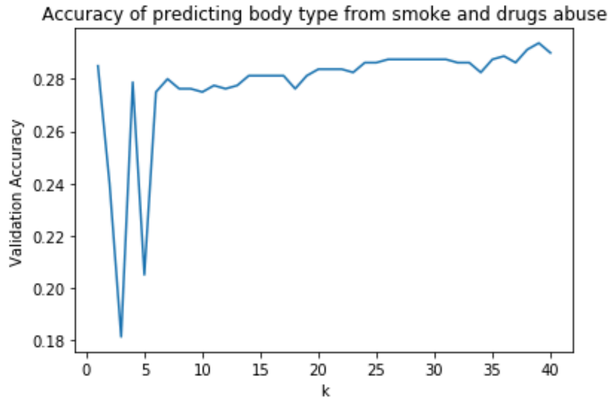
Using KNN at k = 21, the score is 0.28375

The results are not optimal, a guess of the body type would be right 16% of trials.

K-nearest Neighbors Classifier

[21]:

```
start_time = time.time()


from sklearn.neighbors import KNeighborsClassifier

accuracies = []
k_list = list(range(1,41))

for k in range(1,41):
    classifier = KNeighborsClassifier(n_neighbors = k)
    classifier.fit(X_train, y_train.ravel())
    accuracies.append(classifier.score(X_test, y_test.ravel()))


plt.plot(k_list, accuracies)
plt.xlabel("k")
plt.ylabel("Validation Accuracy")
plt.title("Accuracy of predicting body type from smoke and drugs abuse")
plt.show()

print("--- %s seconds ---" % (time.time() - start_time))
```




```
--- 1.5406608581542969 seconds ---
```

Precision & Recall

The mapping for y was:

- ‘fit': 1,
- 'thin': 2,
- ‘average': 3,
- 'athletic': 0,’
- curvy': 5,
- 'a little extra': 4

Recall and precision shows a good prediction rate for average, fit and athletic body types

Recall score

```
[0.         0.01666667 0.05633803 0.90163934 0.         0.        ]
```

Precision score

```
[0.         0.08823529 0.2        0.29490617 0.         0.        ]
```

```
classifier.classes_
```

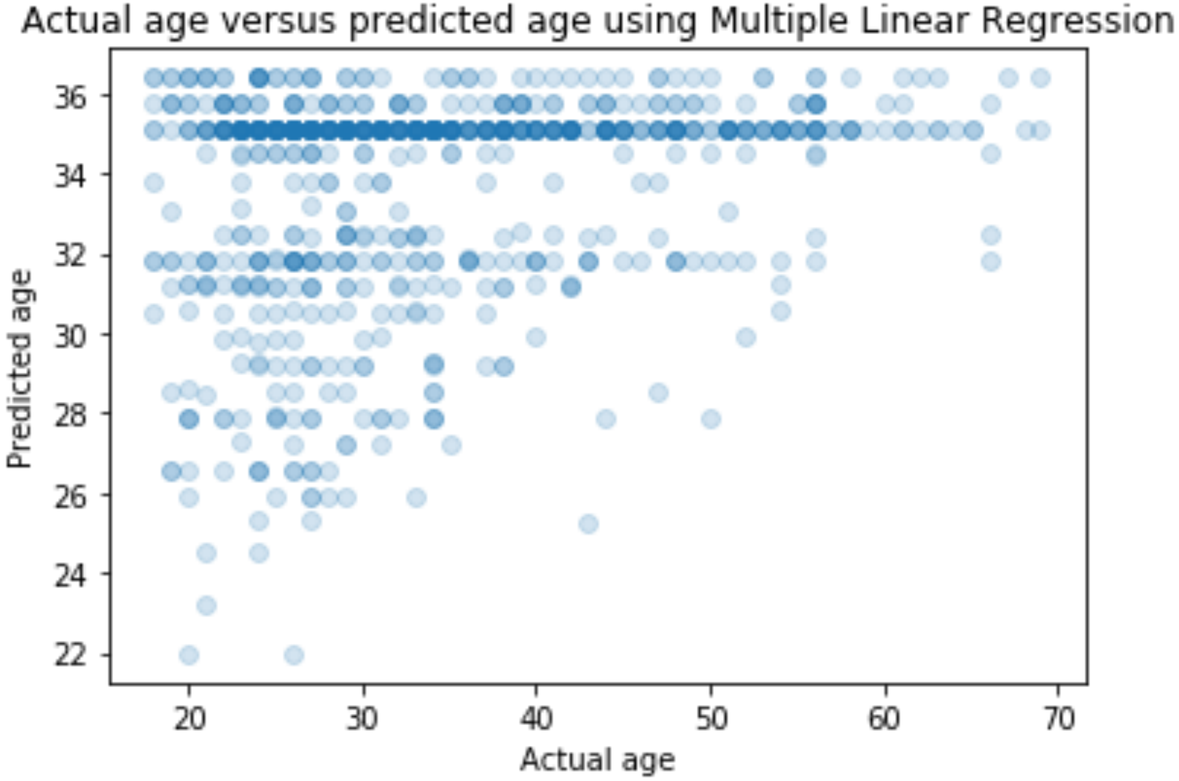
```
array([0, 1, 2, 3, 4, 5])
```

Regression Approaches

Predicting age from drink, smoke and drugs abuse using Linear Regression and K-nearest Neighbours Regressor.

The accuracy for K-nearest Neighbours Regressor is too low to be considered. time to run the mode is 20 seconds.

Both models shows a problem with data. The assumption to be able to predict age on these three features has to be discarded and rethought.



Conclusions/Next steps

I think further value could be extrapolated with an initial cleansing and augmentation of qualitative answers.

My initial hypothesis maybe have been too broad and the results are not very satisfactory. It has been very interesting exploring this real user base, it would be nice to analyse aspects tied to this social network as numbers of matches.