Codecademy Machine Learning Fundamentals Capstone project

Eric Boniface - November 12, 2018

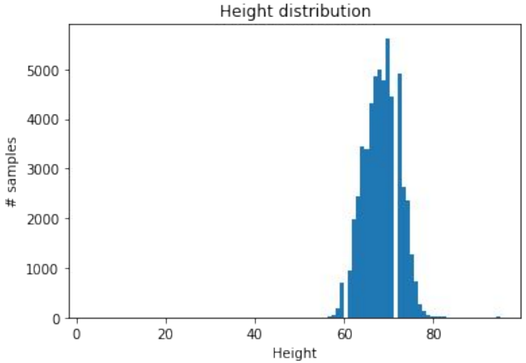
Agenda

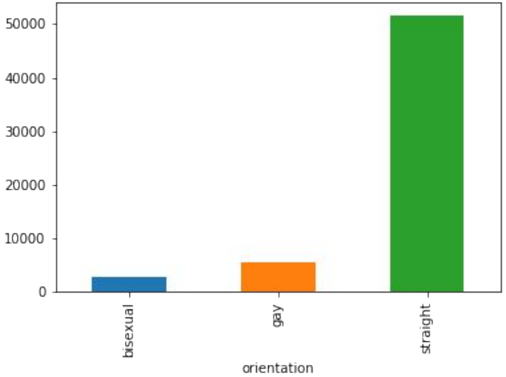
1. Dataset exploration
2. Creation of 2 new features
3. Formulation of a ML question
4. Data preparation
5. Comparison of modelling approaches
6. Conclusion

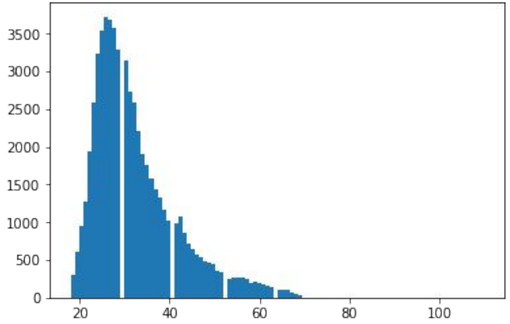
1. Dataset exploration

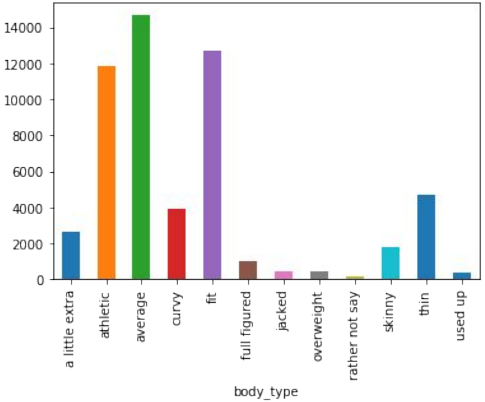
- the dataset contains 60k samples and 31 features
- on some features it is quite sparse:
    - e.g. offspring has 35k n/a, diet has 24k n/a
    - e.g. 13 features have more than 10k n/a
- some features seem to be mandatory fields for the users: age, income, orientation, sex, status have no n/a
    - but users tend to circumvent this, e.g. income has 43k dummy values '-1'

1. Dataset exploration







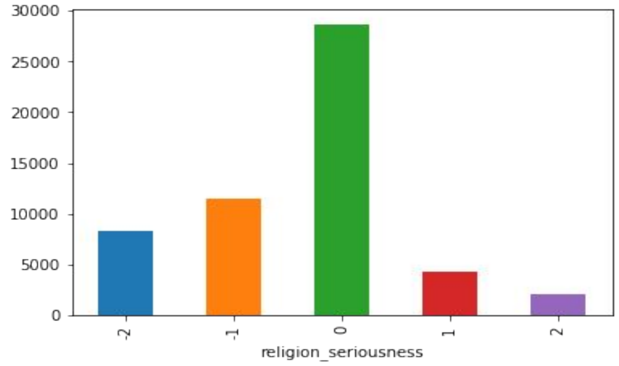


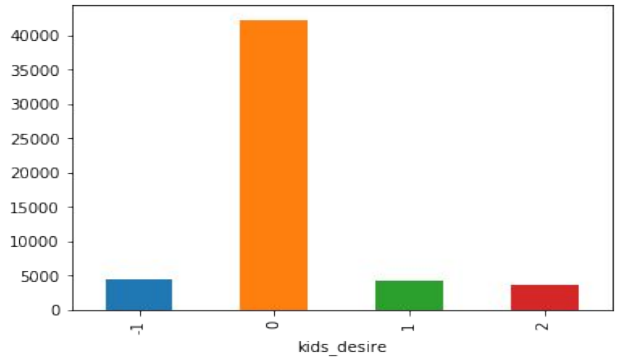
2. Creation of 2 new features

We create 2 new features, extracted from the categorical features 'religion' and 'offspring'. Beyond the base options they describe, these 2 variables contain additional information about:

- the seriousness of the religion involvement, across religions
- the desire to have kids, whether having kids already or not

We extract this information based on the text indication of the level of religion seriousness (resp. kids desire). We consider that no indication of religion seriousness means an average level of seriousness about it, and no indication of kids desire means that the person doesn't know if it wants kids or not. The enables ordering the levels of religion seriousness (resp. kids desire)





3. Formulation on an ML question

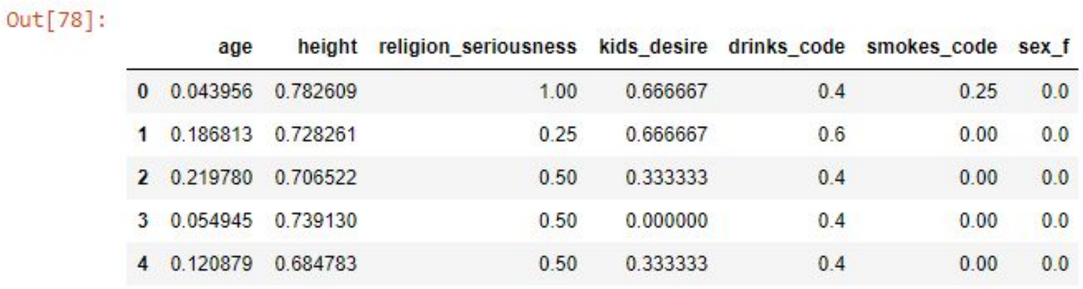
The body type seems like an interesting characteristic to study. In dating situation it contributes significantly to the first impression the two persons have from each other. Some people work actively on shaping their bodies, other suffer it. There are a lot of stereotypes and clichés about body types / shapes, but on the other hand there are also real correlations that exist.

We are wondering if we could manage to predict it from other descriptive features like:

- age, height, sex: physiological descriptors
- drinks, smokes: behavioral descriptors
- religion_seriousness, kids_desire: opinion/mindset descriptors

4. Data preparation

1. We dropped the rows with body_type n/a values
2. We filled drinks and smokes n/a values with drinks and smokes modes
3. We mapped body_type, drinks, smokes to numerical values
4. We mapped sex to dummy values, and dropped one of the 2 resulting redundant columns
5. We normalized all feature values with MinMaxScaler



Out[78]:

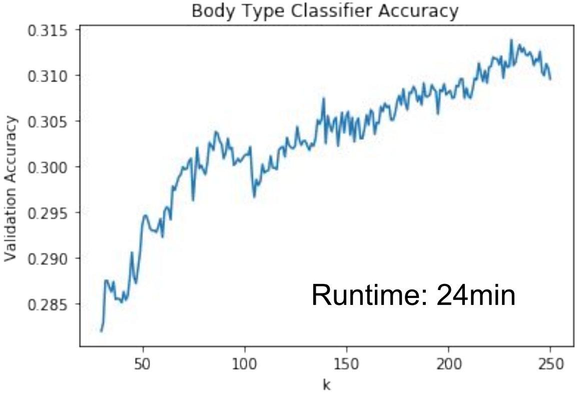
	age	height	religion_seriousness	kids_desire	drinks_code	smokes_code	sex_f
0	0.043956	0.782609	1.00	0.666667	0.4	0.25	0.0
1	0.186813	0.728261	0.25	0.666667	0.6	0.00	0.0
2	0.219780	0.706522	0.50	0.333333	0.4	0.00	0.0
3	0.054945	0.739130	0.50	0.000000	0.4	0.00	0.0
4	0.120879	0.684783	0.50	0.333333	0.4	0.00	0.0

5. Comparison of modelling approaches Classifiers

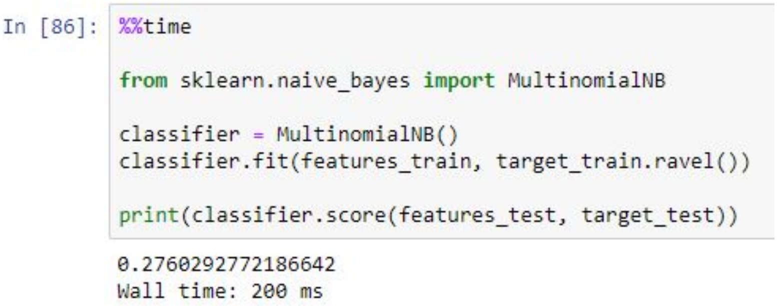
We compared K-Nearest Neighbors and Naive Bayes Classifier approaches.

The accuracy we reach is very low, 31,4% with K-Nearest Neighbors vs. 27,6% with Naive Bayes Classifier. But it is still fairly above random, since with 11 classes random would be 9%.

Finding the best k in the K-Nearest Neighbors approach took approx. 30min of computing on a standard laptop, including some trials and errors; whereas the Naive Bayes Classifier approach is much simpler.



Runtime: 24min



```
%%time

from sklearn.naive_bayes import MultinomialNB

classifier = MultinomialNB()
classifier.fit(features_train, target_train.ravel())

print(classifier.score(features_test, target_test))
```

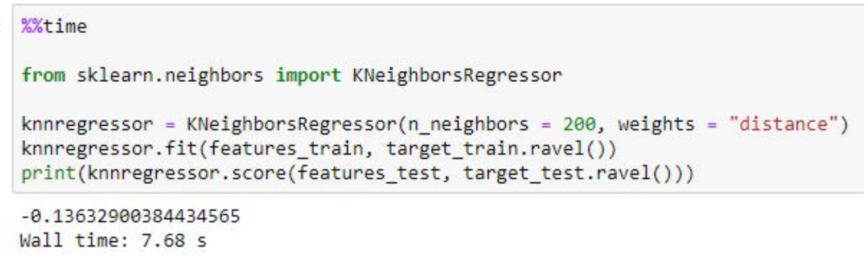
```
0.2760292772186642
Wall time: 200 ms
```

5. Comparison of modelling approaches Regressors

We compared K-Neighbors Regressor and Multiple Linear Regression approaches.

The performance we obtained is disastrous.

This could indicate a problem either in our training data or in our implementation of the regressors.

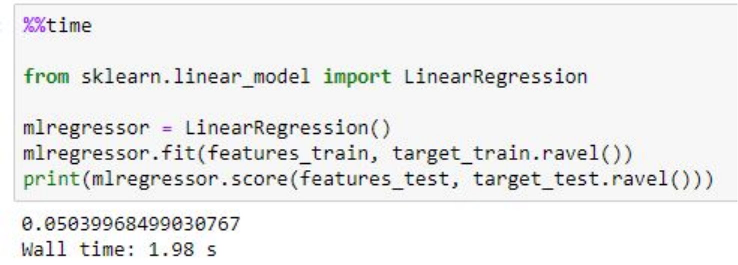


```
%%time

from sklearn.neighbors import KNeighborsRegressor

knnregressor = KNeighborsRegressor(n_neighbors = 200, weights = "distance")
knnregressor.fit(features_train, target_train.ravel())
print(knnregressor.score(features_test, target_test.ravel()))
```

```
-0.13632900384434565
Wall time: 7.68 s
```



```
%%time

from sklearn.linear_model import LinearRegression

mlregressor = LinearRegression()
mlregressor.fit(features_train, target_train.ravel())
print(mlregressor.score(features_test, target_test.ravel()))
```

```
0.05039968499030767
Wall time: 1.98 s
```

6. Conclusion

We experimented predicting the body type from:

- age, height, sex: physiological descriptors
- drinks, smokes: behavioral descriptors
- religion_seriousness, kids_desire: opinion/mindset descriptors

The experiment is a failure. Our initial intuition was wrong, there seem to be no correlation between body type and these descriptors.

Comments on the classification approaches:

- it seems delicate to predict a target with 11 different possible classes, of which some are very close

Comments on the regression approaches:

- our target data might not have been super significant: the mapping of the body type classes to ordered numerical values is quite uncertain