Machine Learning Fundamentals

Analysis of OKCupid data

Birger Kollstrand , 11.Nov.2018

Table of Contens

- Exploration of the Dataset
- Question to Answer
- Augmenting the Dataset
- Approaches
- Conclusions/Next steps

Exploration of the Dataset

The dataset contains 59946 rows and 31 columns of varied data.

The columns in the dataset are 'age', 'body_type', 'diet', 'drinks', 'drugs', 'education', 'essay0', 'essay1', 'essay2', 'essay3', 'essay4', 'essay5', 'essay6', 'essay7', 'essay8', 'essay9', 'ethnicity', 'height', 'income', 'job', 'last_online', 'location', 'offspring', 'orientation', 'pets', 'religion', 'sex', 'sign', 'smokes', 'speaks' and 'status'

The data set has been reviewed in the dating_skeleton_review.py file.

The review highlighted that the data set is complex and with varied type of data.

Question to Answer

Question:

Is there a relevance between the users smoking, drinking, drug use, height how much they write about them selves and their sex?

Agumentation of the Dataset

Several data are not in a format usable for Machine learning.

Some data was mapped to numerical values and expanding the number of columns. The new columns are prefixed with “cust_”

'age', 'body_type', 'cust_bodyType', 'diet', 'cust_diet', 'drinks', 'cust_drinks', 'drugs', 'cust_drugs', 'education', 'essay0', 'essay1', 'essay2', 'essay3', 'essay4', 'essay5', 'essay6', 'essay7', 'essay8', 'essay9', 'essay_len', 'ethnicity', 'height', 'income', 'job', 'last_online', 'location', 'offspring', 'orientation', 'cust_orientation', 'pets', 'religion', 'sex', 'cust_sex', 'sign', 'smokes', 'cust_smokes', 'speaks', 'status', 'cust_status'

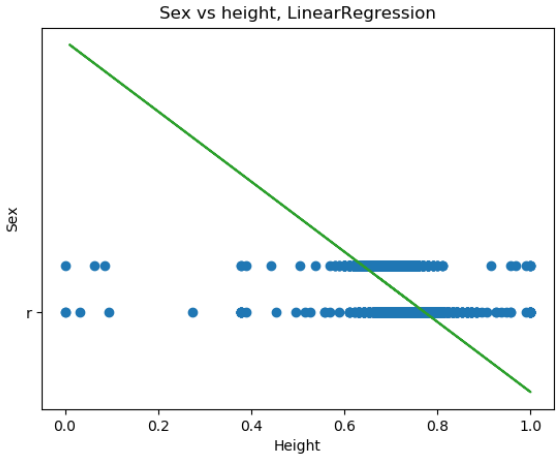
Details of the customization can be reviewed in the dating_skeleton_prep_data.py file. The modified data is written to a new CSV file to avoid unwanted changes to the original data set.

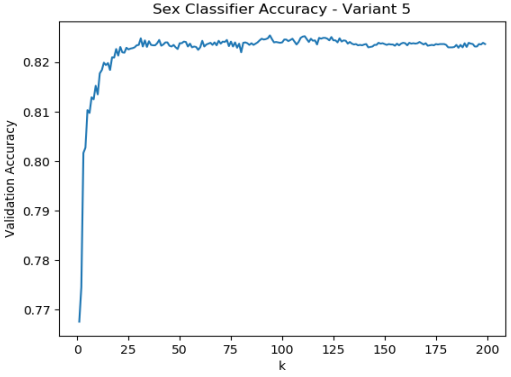
Approaches

- Looked at linear regression classification
- And KNeighbours classification

The LinearRegressions run with different combinations all yielded a very low score. The with Height vs Sex at 0.42.

KneighboursClassifier was then used in a number of variants.

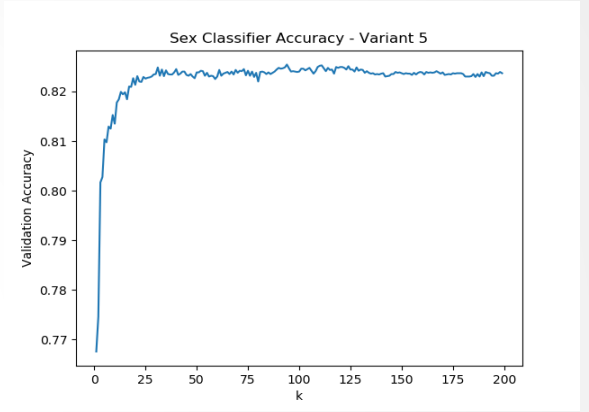




KNeighboursClassification

- 6 variants tested, score is before the parameter list

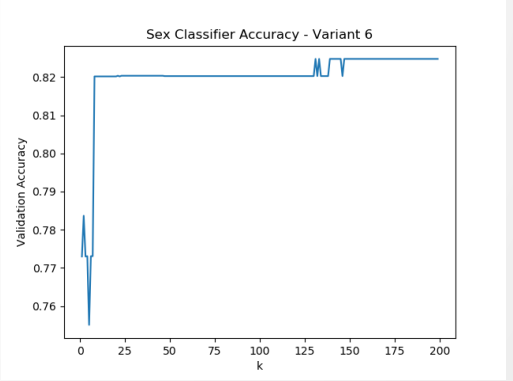
1) 0.8178 k=39 ['cust_smokes', 'cust_drinks', 'cust_drugs', 'height', 'essay_len']
2) 0.5939 k=39 ['cust_smokes', 'cust_drinks', 'cust_drugs']]
3) 0.8213 k=8 ['cust_smokes', 'cust_drinks', 'cust_drugs', 'height']
4) 0.5906 k=184 ['cust_smokes', 'cust_drinks', 'cust_drugs', 'essay_len']
5) 0.8254 k=94 ['height', 'essay_len']
6) 0.8248 k=131 ['height']



The LinearRegressions run with different combinations all yielded a very low score. The with Height vs Sex at 0.42.

KneighboursClassifier was then used in a number of variants. All variants tested had a better score than I achieved with linear regression.

Variant 5 and 6 yields very similar results, but the calculation time for variant 6 was significantly shorter than for variant 5.



Conclusions/Next steps

Using only parameters in regards to intoxication is not relevant. Also the length of the essays seems to be irrelevant to the classification result.

Including “Height” does significantly increase the score, but it also works equally well to only use height.

It is not accurate enough to find male/female based upon the parameters selected.

Based the this work I would not recommend to use this without further analysis.