„Date a scientist“

Capstone Project for Codeacademy Course
Machine learning Fundamentals
Jakob Theileis
12.11.2018

Table of content

Dataset

- Dataset of OK-Cupid from 2011 – 2012 provided by codeacademy
- 59946 rows
- 31 columns ('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', 'status')

Question to be asked

- Classification

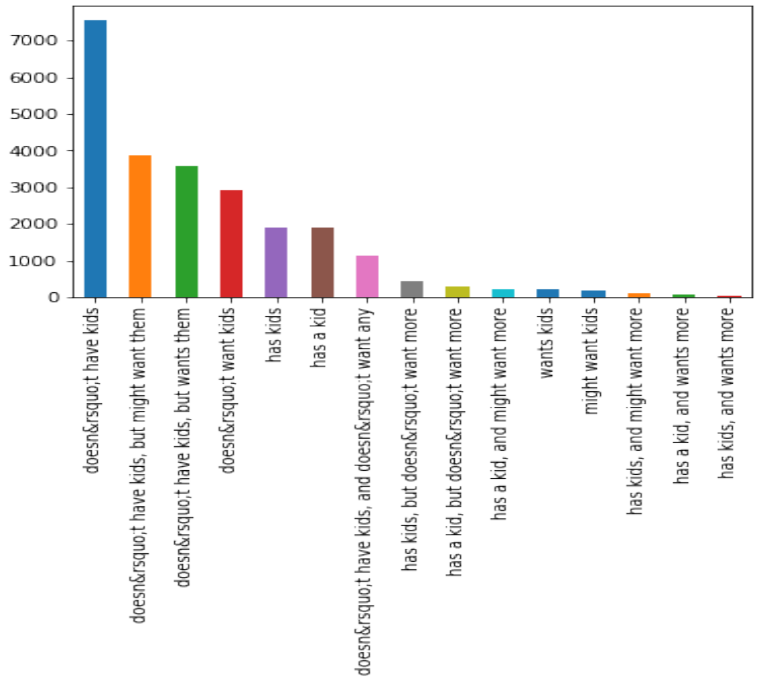
  Can you predict the offspring with given dataset?

- Regression

  Can you predict the age of a person?

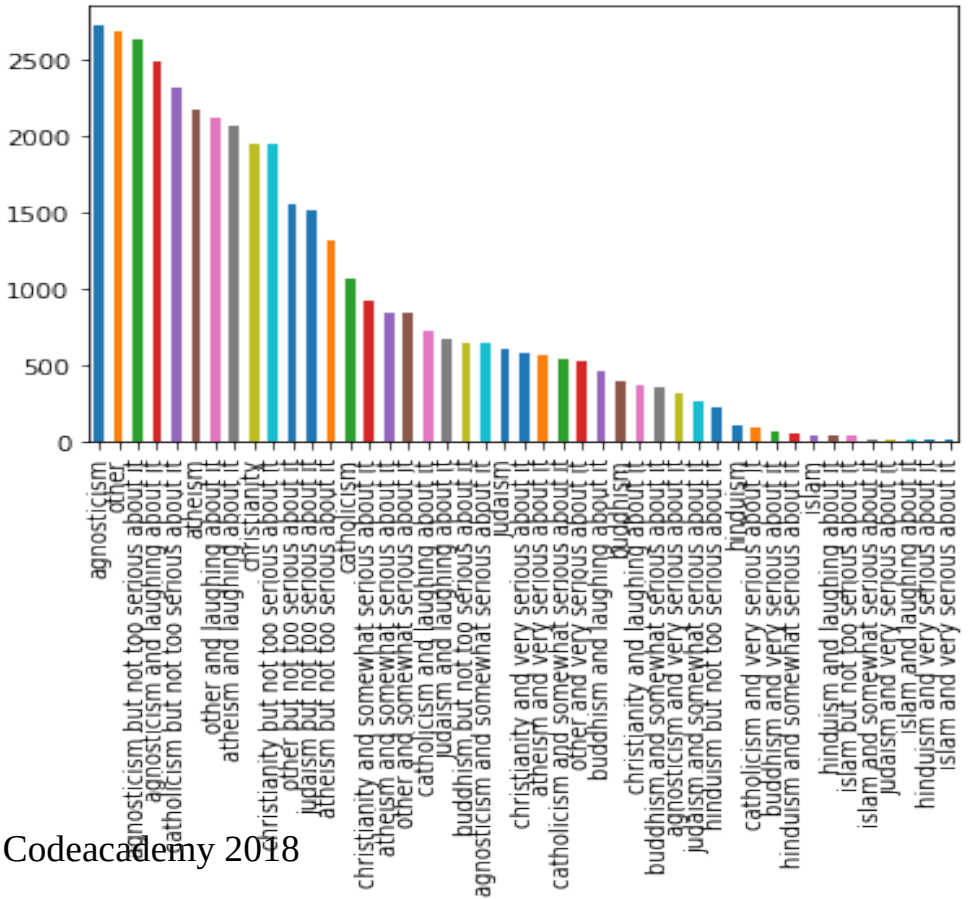
Exploration of data

- Using .value_counts() func to look at column offspring
- High amount of people who „doesn‘t have kids“
- Compliant with data expected from dating website



Exploration of data

- Using .value_counts() func to look at column **religion**
- Lot of different religions
- Different categories of religious engagment from ‚laughing about it‘ to ‚very serious about it‘



Augmentation

Mapping

- Categorical Data → Numerical Data

  1) Automatical Mapping (AM) function uses value counts to transform data:

  (ie. ‚doesn‘t have kids ‘:0 … ‚has kids and wants more kids‘:15

  2)Hand mapping (HM) trying to give the values a linear form

  (ie. 'agnosticism and laughing about it': 0 … 'islam and very serious about it': 4

- Can Automatical Mapping be used in this case?

Evaluation

Hand Mapping vs Automatic Mapping

- Automatic mapping has in this case a slightly better recall score
- AM faster to evaluate more data
- Both mapping methods have a low score

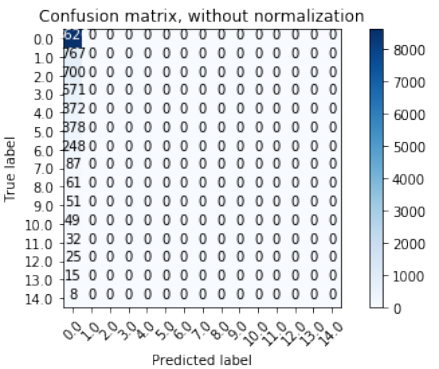
	Hand Mapping	Automatic Mapping
Features (normalized)	Religion_lin, Status_lin	Religion, Status
Labels	Offspring	Offspring
Recall Score (KNN-Class. k=55)	0.3191395961 3696224	0.325504828 7971905
Recall Score (MNB-Class.)	0.3154082528 5338014	0.315627743 63476737

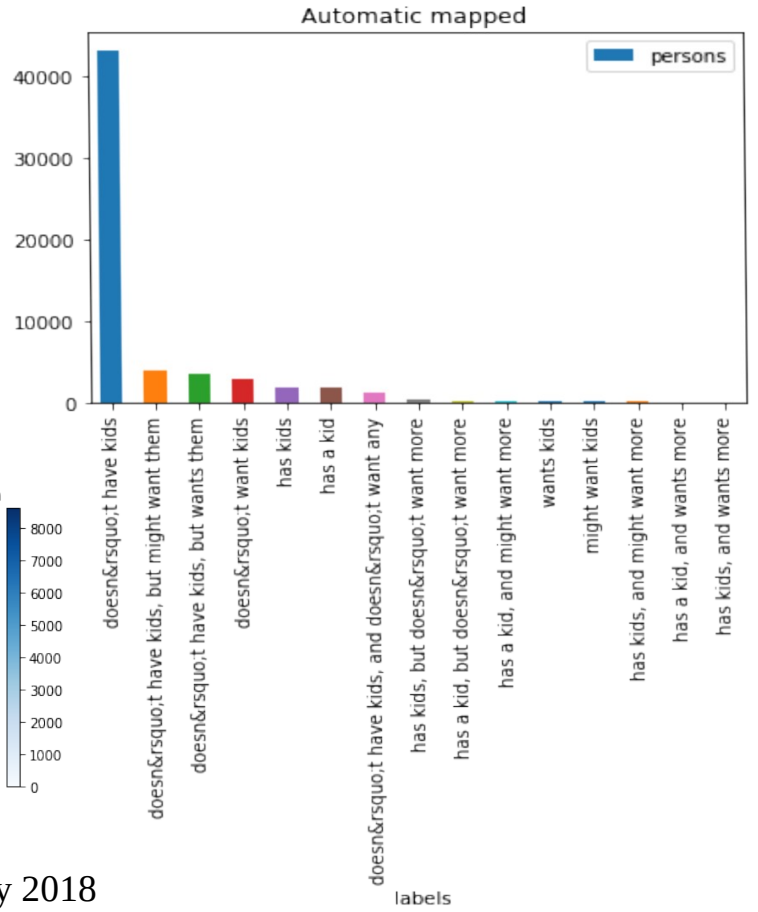
Problem of imbalanced classes

- By using *df.fillna({'offspring_code':0,}, inplace=True)* one class will be overrepresented → imbalanced dataset
- Results in a „false good“ score (ie. KNN Class k=55. Score: 0.7194)
- „false good“ Score is because of one true overrepresented class(ie. ‚doesn‘t want kids‘)
- Distribution of trues in con. matrix very bad
- Solution:

  - drop rows containing NaN (loss of data in order get better results)

  - Modifying classes

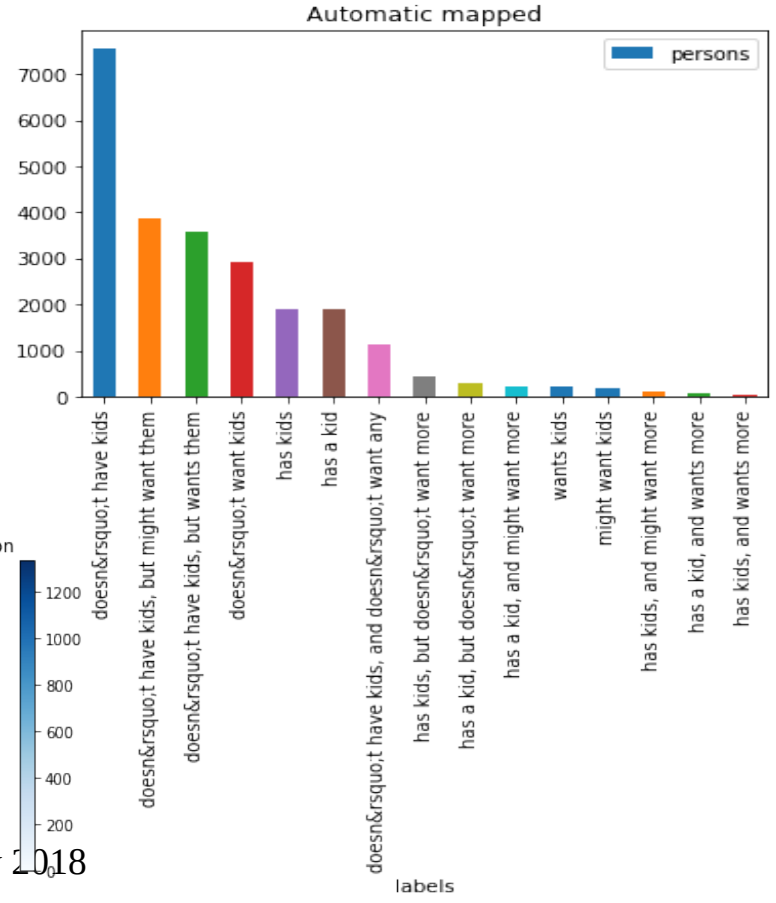


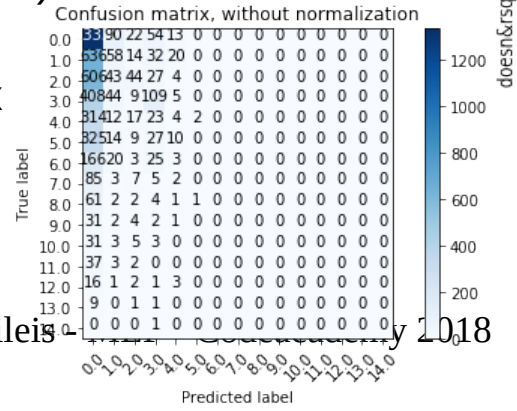


Problem of imbalanced classes

dropping rows with nan

- Dropping 35561 rows with nan (loss of information)
- Dataset still contains 24385 entries.
- For now we use the dataset without the dropped rows
- For future examination we should take a closer look at the dropped rows
- Score dropped to 0.3174 (KNN k=55 featuredata='religion_code', 'status_code')
- Still one class overrepresented but distribution of trues better in con. matrix
- Next step: modifiyng classes





Problem of imbalanced classes

modifying classes

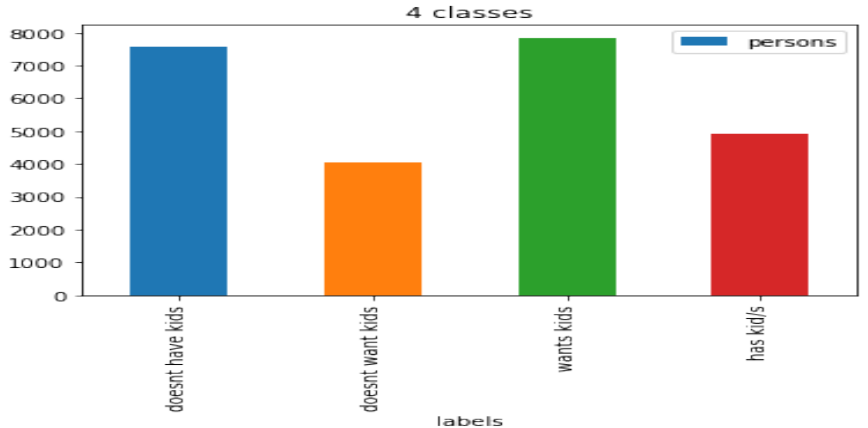
- I rearranged the dataset in 4 classes

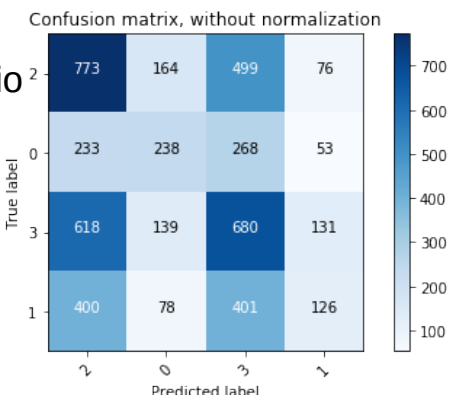
  1: doesnt have kids (neutral)

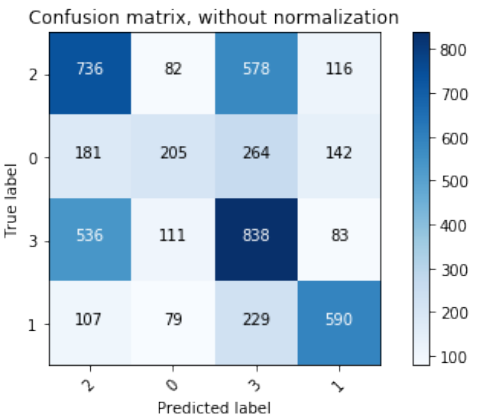
  2: doesn‘t want kids

  3: wants kids but has no kids

  4: has kid/s
- Score raised to 0.3726 (KNN k=55)
- Better distribution of trues in confusion matrix
- To improve the score i used more feature data ('religion_code','income','education_code', 'status_code','age','ethnicity_code','sex_code','orientation_code','pets_code')
- Score raised to 0.4857 (KNN k=55)
- Distribution of trues better in conf. matrix

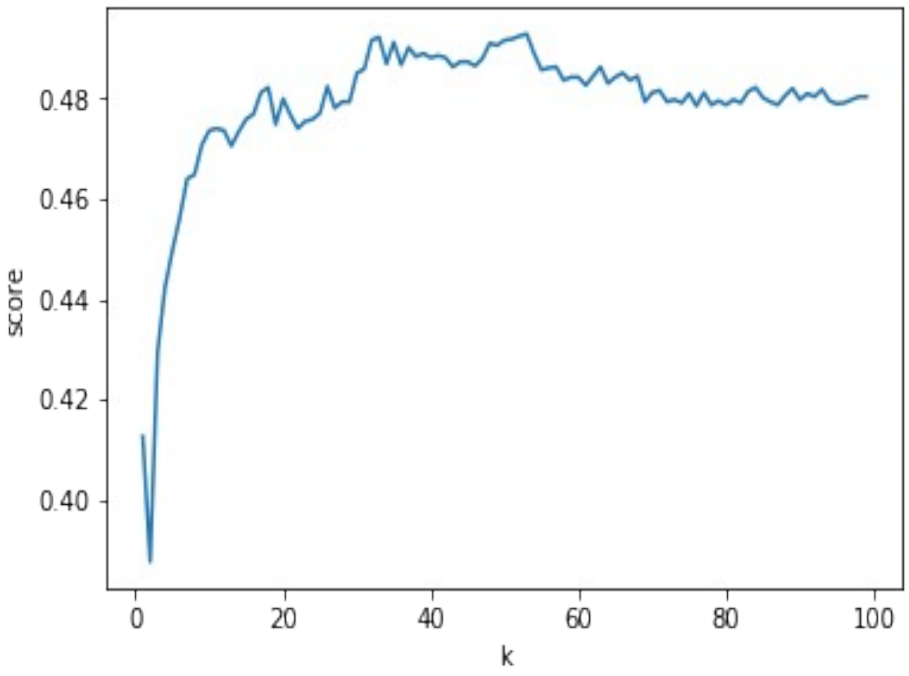






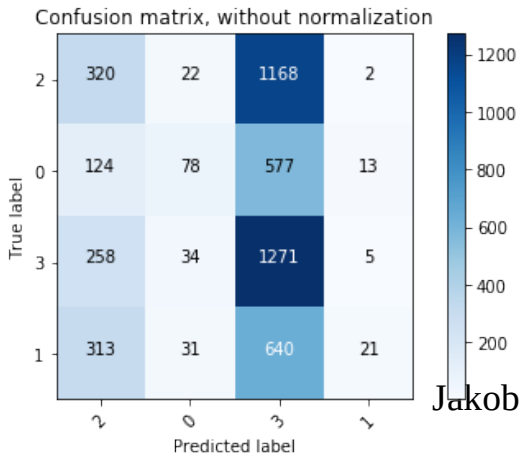
K Nearest Neighbors Classifier

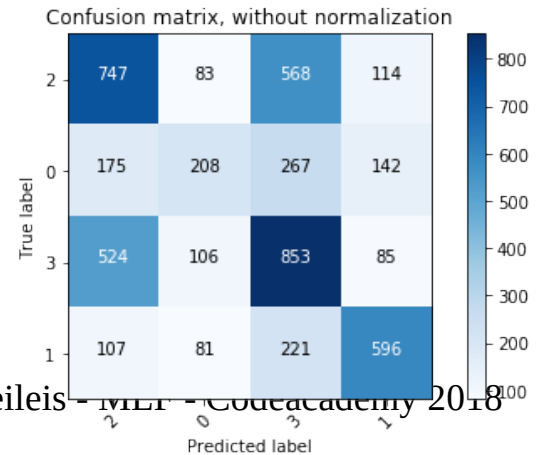
- In order to find the best k I looped through 1:100K
- Best k = 53
- Best score = 0.4929

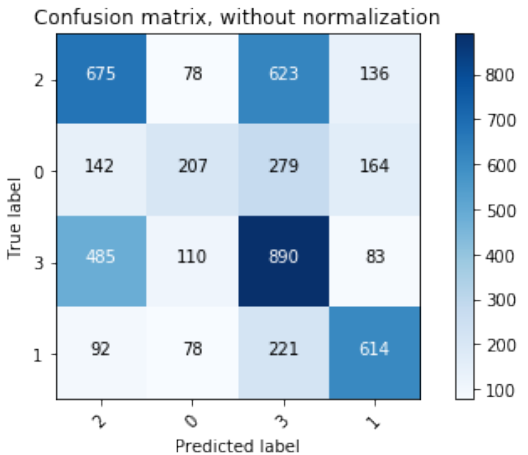


Evaluation of Classification models

	Naives Bayes Classifier	K Nearest Neighbors k=53	Support Vector c=7 gamma=1
Accuracy	0.3465	0.4929	0.4892
Recall	0.3465	0.4929	0.4892
F1	0.2697	0.4886	0.4833
Time[s]	0.0210	0.8380	25.4914
simplicity	easy	modest	modest







Regression

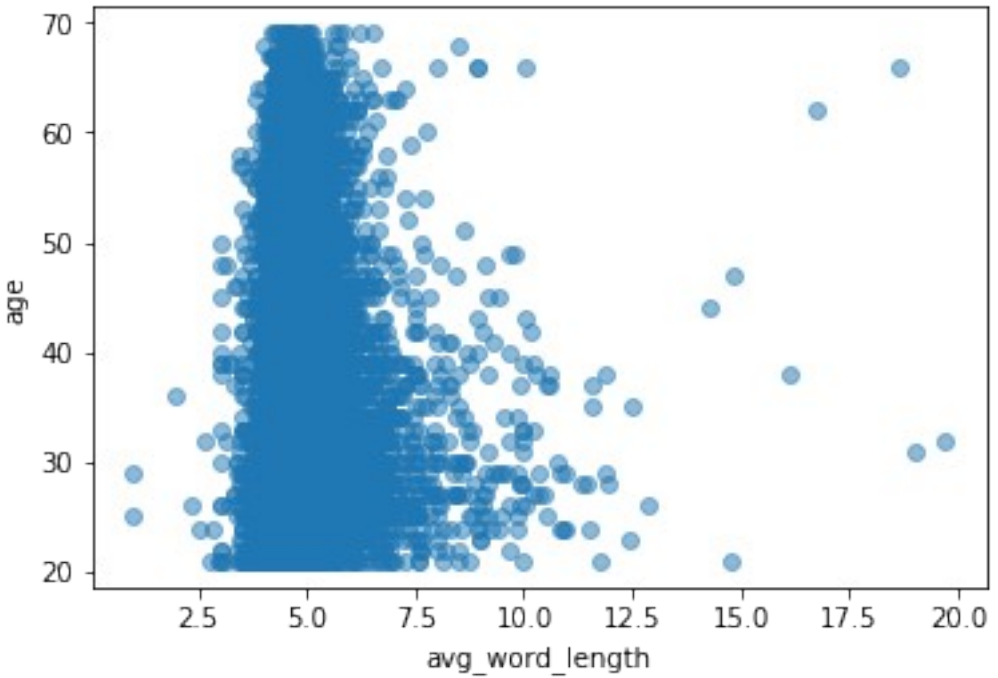
- In order to to examine regression I created new columns:

column	content
df["avg_word_length"]	average length of words in essays
df["essay_I"]	number of times ‘I‘ appears in essays (same for ‘me‘ and ‘you‘)
df["speaks_number"]	number of languages a person speaks

Cols I use to predict age	['avg_word_length','essay_len','education_code','speaks_number','income','offspring_code','word_unique_count']

Preparing the data

- Removing outliers
- Age > 100 :dropped
- Avg word length >25 dropped
- Age <21: dropped



Evaluation of Regression models

	Multiple Linear Regression	K Neighbors Regression
R2	0.1180	0.1194
Mean Absolute Error	7.6864	7.3933
Time[s]	0.0400	0.2510
simplicity	easy	easy

Conclusion

	Classification	Regression
Question	Predicting offspring	Predicting age
Awnser	High chance that you can predict the right awnser if you have no kid. If you have a kid the classifier doesnt work as well. (Imbalanced Data)	Its very unlikely to predict the correct age with regression.
Best Model	K nearest Neighbors Had best results with a decent speed	K nearest Regression Not significantly better results then MLR
Next steps	Closer look in data that has been dropped	Try other data
Data	Offspring is a tricky question for a dating plattform, since peole are mostly singles and singles have less kids.	More numerical data would be good and anyhow how many people are lying about their age while dating :)