OkCupid

Data Science project

Data set

- The project has one data set provided by Codecademy called profiles.csv.
- Each row represents an OkCupid user and the columns are the responses to their user profiles which include multi-choice and short answer questions

	age	body_type	diet	drinks	drugs	education	essay0	essay1	essay2	essay3	...	location	offspring	orientation	pets
0	22	a little extra	strictly anything	socially	never	working on college/university	about me: \n \ni would love to think...	currently working as an international agent fo...	making people laugh. \nranting about a go...	the way i look. i am a six foot half asian, ha...	...	south san francisco, california	doesn’t have kids, but might want them	straight	likes dogs and likes cats
1	35	average	mostly other	often	sometimes	working on space camp	i am a chef: this is what that means. \n1...	dedicating everyday to being an unbelievable b...	being silly. having ridiculous amonts of fun w...	NaN	...	oakland, california	doesn’t have kids, but might want them	straight	likes dogs and likes cats
2	38	thin	anything	socially	NaN	graduated from masters program	i'm not ashamed of much, but writing public te...	i make nerdy software for musicians, artists, ...	improvising in different contexts. alternating...	my large jaw and large glasses are the physica...	...	san francisco, california	NaN	straight	has cats
3	23	thin	vegetarian	socially	NaN	working on college/university	i work in a library and go to school. . .	reading things written by old dead people	playing synthesizers and organizing books acco...	socially awkward but i do my best	...	berkeley, california	doesn’t want kids	straight	likes cats

Project goals

In this project, the goal is to utilize the skills learned through Codecademy and apply machine learning techniques to a data set. The following questions will be addressed in this study:

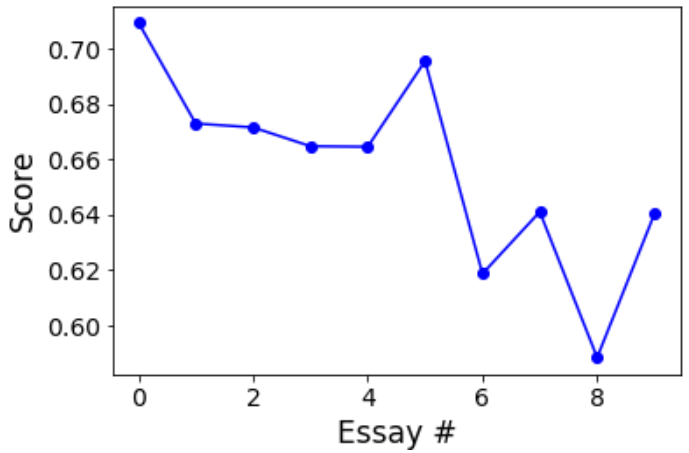
- 1. Can the open answers predict the sex of the user?
- 2. Can a user's astrological sign be predicted from other features of the user?
- 3. Bonus: is there a wage difference between the male and female users?

Q1. Can the open answers predict the sex of the user?

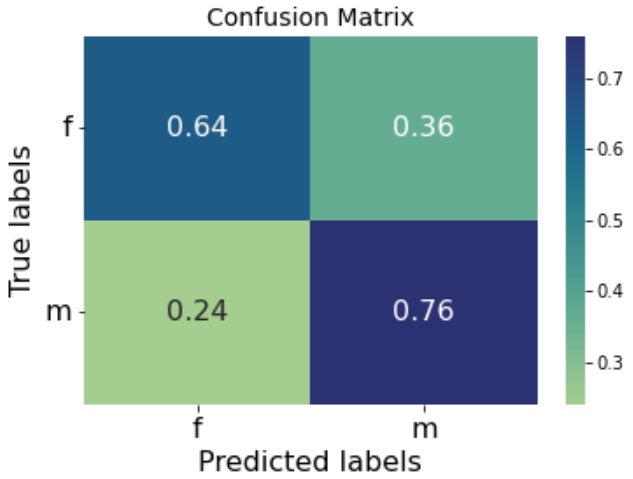
Example open answer

„I am a chef: this is what that means. 1. i am a workaholic. 2. i love to cook regardless of whether i am at work. 3. i love to drink and eat foods that are probably really bad for me. 4. i love being around people that resemble line 1-3. i love the outdoors and i am an avid skier. if its snowing i will be in tahoe at the very least. i am a very confident and friendly. i'm not interested in acting or being a typical guy. i have no time or patience for rediculous acts of territorial pissing. overall i am a very likable easygoing individual. i am very adventurous and\nalways looking forward to doing new things and hopefully sharing it with the right person."

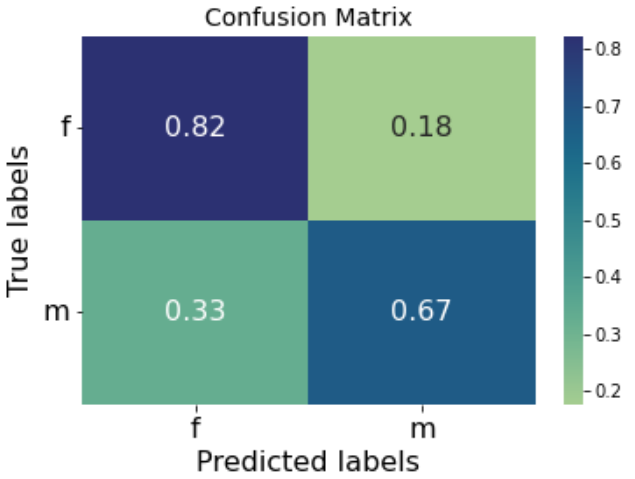
How accurately does each essay predict user‘s sex?



Can the first essay predict the user‘s sex?



Can all 10 essays combined predict better the user‘s sex?

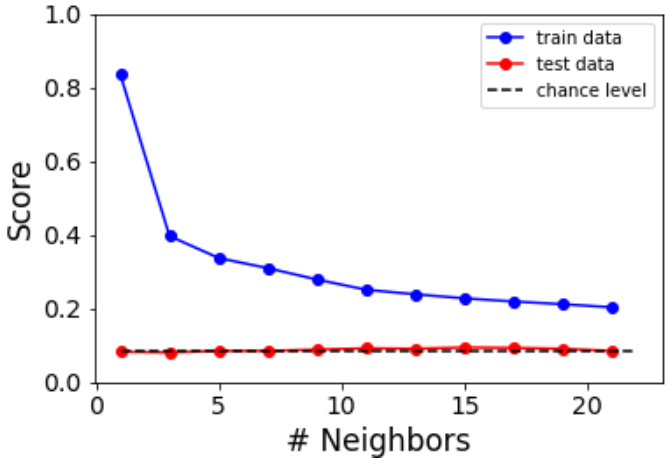


Q2. Can a user's astrological sign be predicted from other features of the user?

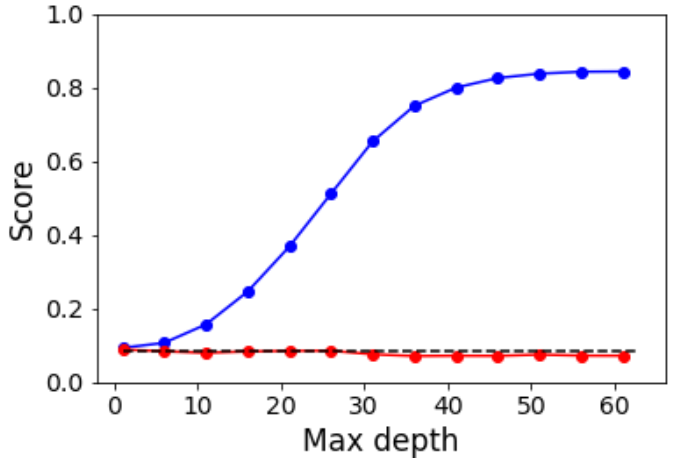
A subset of the data is created, with features which might predict the astrological sign: body type, diet, drinks, drugs, smokes, orientation, has pets, religion, sex, job.

Performance of different models with varying parameters:

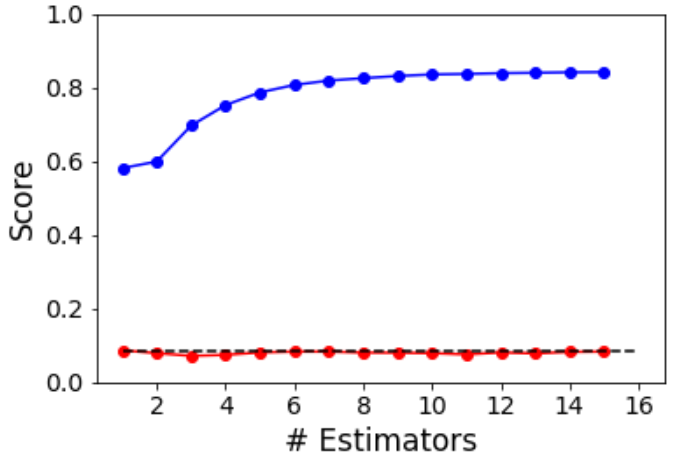
K-nearest neighbors



Decision trees



Random forest



Q3 (bonus): Is there a wage difference between the male and female users?

The histogram and box plot show there is a difference in the income of male vs. female users. The statistical significance is confirmed by the t test ($p < 0.001$).

Median salary earned by men: 60000 $

Median salary earned by women: 40000 $