codecademy

OKCupid – Date –A- Scientist

Cletus Norton
Nov. 28 2020

1.1 Information about the data

This is a capstone project for the data science path, where we are to analyze data from an on-line dating application called OKCupid. As phones have become more computer-like there has been a rise in dating apps to find love. Many of these apps use data science techniques to recommend possible matches in hope to connect users to their “true Love”. The amount of data collected via the app gives us a large amount of personal information about human romance.

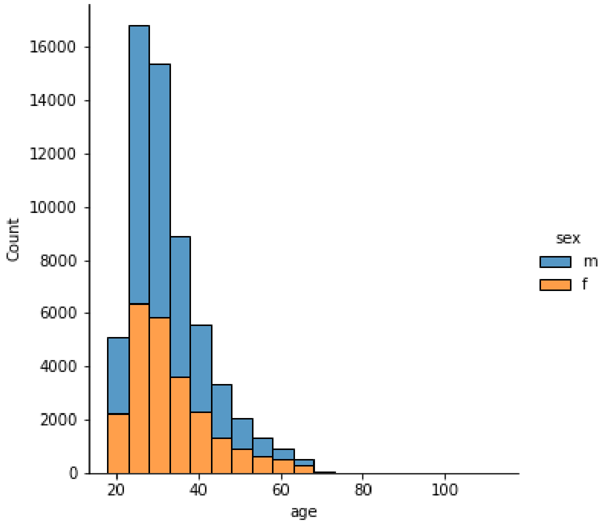
The project goal is to try and perdict the astrological sign of the OKCupid users by using the other varibles provided. This is a fun use of the data since many users find astrological signs as a fun aspect of dating. Since the data provided doesn’t include any birth imformation beyond age, we developed a means to predict the correct sign.

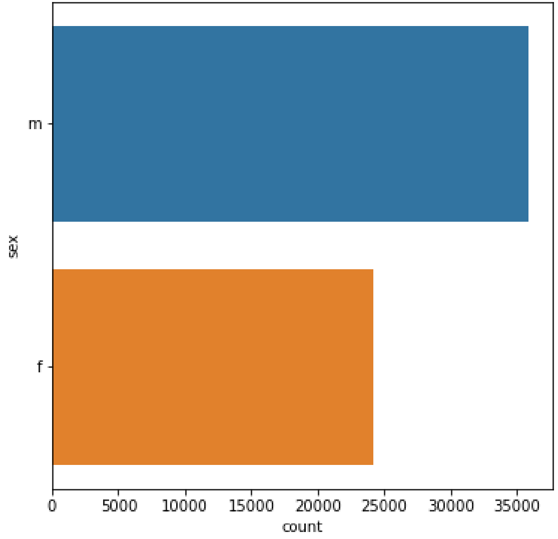
Using the data provided in “profiles.csv”, I present a solution that will use descriptive statistics and data vizualization to find important features to understand the counts, relationship and distribution between the features. Supervised learning models of machine learning were used to make the predictions. Final analysis of the predictions will be checked via accuracy metrics and confusion matix.

1.2 Dataset detail

- Count of the dataset of men and women
- Compares age between the sexes
- Most of both groups are in mid-twenties to late thirties

- Count showing the number of men versus women
- There are about 10,000 more men in the dataset

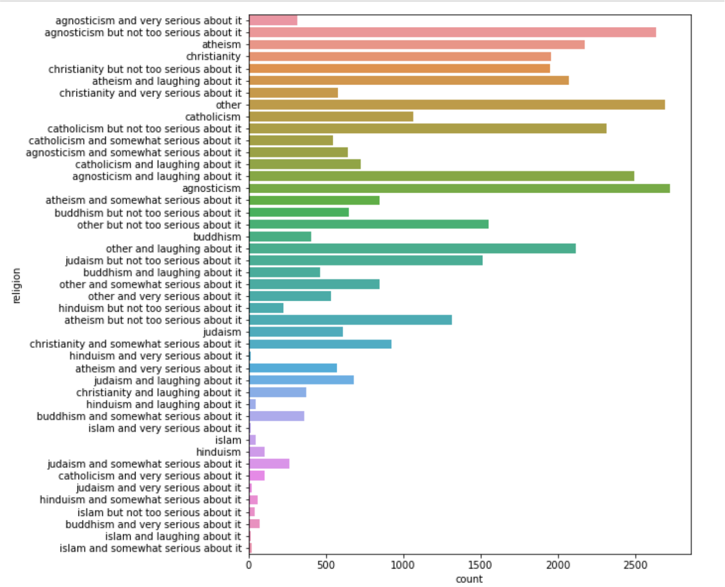


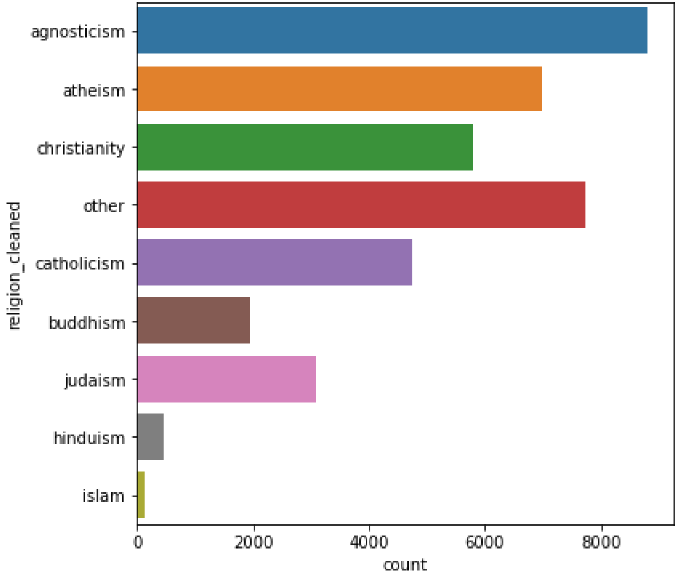


1.3 Dataset detail

- Religion responses
- Several had comments about strength of belief

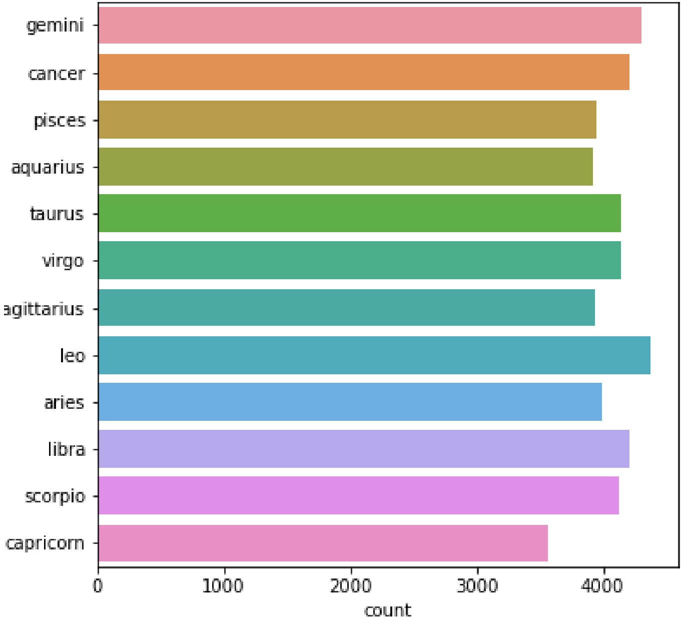
- Count of religion response after clean up
- Comments about strength of belief were removed





1.4 Astrological signs

- This show the counts of the astrological signs after removing comments about interest/belief
- This shows that we have a proportional count on each sign



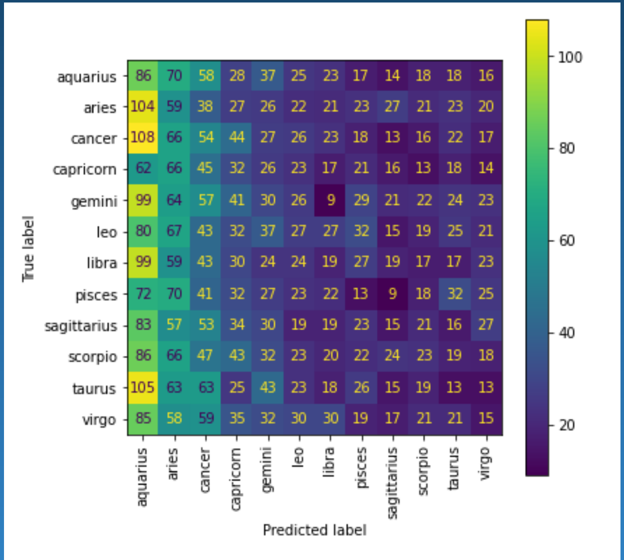
In my effort to determine the astrological sign based on variables provided, I tried two supervised machine learning approaches. The first was the K-nearest neighbor, while the second was using skLearn’s DecisionTree. In the next couple of slides I will go over the details of how each faired. After looking over the variables, I chose the following to use in my machine learning models:

Body Type, diet, orientation, education, religion, sex, & job.

I chose to leave, drinking, drug use, ethnicity, height, income, job, location, offspring, pets, smokes, language spoken, & the essay answers out of the model.

1.5 K-Nearest Neighbor

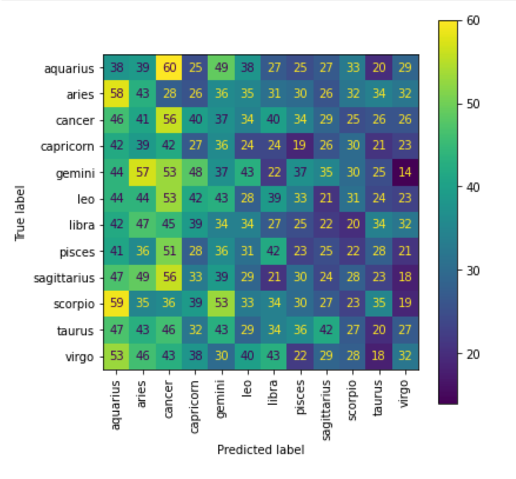
- The data was split into 75% train and 25% test.
- The model showed a mean accuracy of 32%.
- The confusion matrix plot below shows a rather uneven success rate.



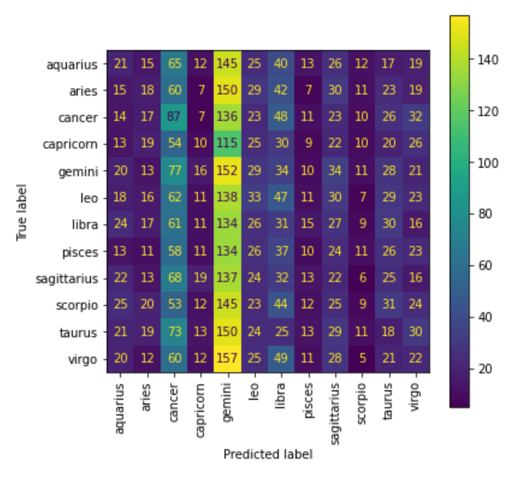
	aquarius	aries	cancer	capricorn	gemini	leo	libra	pisces	sagittarius	scorpio	taurus	virgo
aquarius	86	70	58	28	37	25	23	17	14	18	18	16
aries	104	59	38	27	26	22	21	23	27	21	23	20
cancer	108	66	54	44	27	26	23	18	13	16	22	17
capricorn	62	66	45	32	26	23	17	21	16	13	18	14
gemini	99	64	57	41	30	26	9	29	21	22	24	23
leo	80	67	43	32	37	27	27	32	15	19	25	21
libra	99	59	43	30	24	24	19	27	19	17	17	23
pisces	72	70	41	32	27	23	22	13	9	18	32	25
sagittarius	83	57	53	34	30	19	19	23	15	21	16	27
scorpio	86	66	47	43	32	23	20	22	24	23	19	18
taurus	105	63	63	25	43	23	18	26	15	19	13	13
virgo	85	58	59	35	32	30	30	19	17	21	21	15

1.6 DecisionTree model

- The chart on the left shows the decision-tree model used with no depth limit, resulting in a mean accuracy of 73%.
- The chart on the right shows the decision-tree model used with a depth limit of 20, resulting in a mean accuracy of 28%.
- Without limiting the tree depth (which originally was 64) the decision tree overfit the data.
- When we begin to limit the depth the accuracy dropped below the KNN method.



True label	aquarius	aries	cancer	capricorn	gemini	leo	libra	pisces	sagittarius	scorpio	taurus	virgo
aquarius	38	39	60	25	49	38	27	25	27	33	20	29
aries	58	43	28	26	36	35	31	30	26	32	34	32
cancer	46	41	56	40	37	34	40	34	29	25	26	26
capricorn	42	39	42	27	36	24	24	19	26	30	21	23
gemini	44	57	53	48	37	43	22	37	35	30	25	14
leo	44	44	53	42	43	28	39	33	21	31	24	23
libra	42	47	45	39	34	34	27	25	22	20	34	32
pisces	41	36	51	28	36	31	42	23	25	22	28	21
sagittarius	47	49	56	33	39	29	21	30	24	28	23	18
scorpio	59	35	36	39	53	33	34	30	27	23	35	19
taurus	47	43	46	32	43	29	34	36	42	27	20	27
virgo	53	46	43	38	30	40	43	22	29	28	18	32



True label	aquarius	aries	cancer	capricorn	gemini	leo	libra	pisces	sagittarius	scorpio	taurus	virgo
aquarius	21	15	65	12	145	25	40	13	26	12	17	19
aries	15	18	60	7	150	29	42	7	30	11	23	19
cancer	14	17	87	7	136	23	48	11	23	10	26	32
capricorn	13	19	54	10	115	25	30	9	22	10	20	26
gemini	20	13	77	16	152	29	34	10	34	11	28	21
leo	18	16	62	11	138	33	47	11	30	7	29	23
libra	24	17	61	11	134	26	31	15	27	9	30	16
pisces	13	11	58	11	134	26	37	10	24	11	26	23
sagittarius	22	13	68	19	137	24	32	13	22	6	25	16
scorpio	25	20	53	12	145	23	44	12	25	9	31	24
taurus	21	19	73	13	150	24	25	13	29	11	18	30
virgo	20	12	60	12	157	25	49	11	28	5	21	22

In conclusion this was a fun lesson to apply some basic machine learning to, even though the model ended up no better than guessing. There may have been better data in the essay answers but finding a way to quantify them proved outside my current skills.