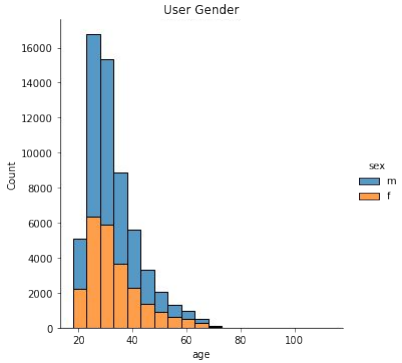
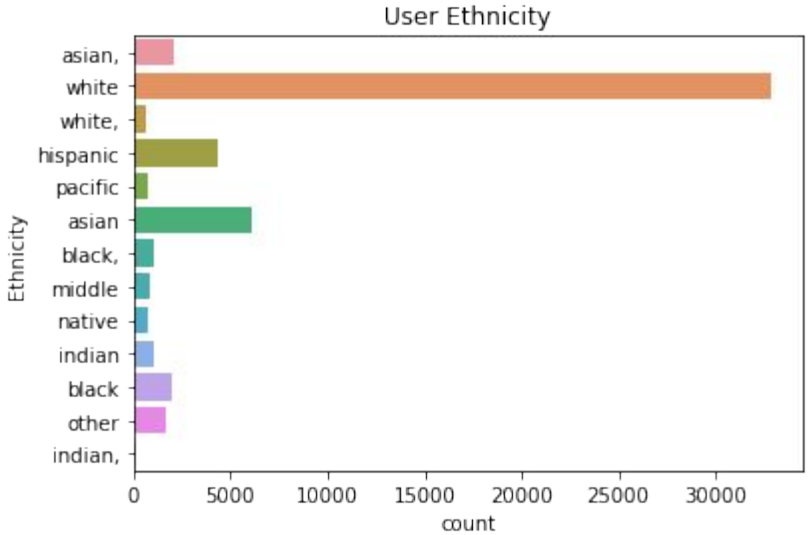
Date-A-Scientist Sign Predictor!

By Dustin Spence

A Look at the Data

- Using data from the popular dating site OkCupid revealed that entries are incredibly inconsistent.
- Many fields are left blank
- Many fields are incredibly skewed in one direction
  - White Men for example make up a very large portion of the data
- It is possible that data points that skew in a certain direction can create trends in the overall analysis of the data.



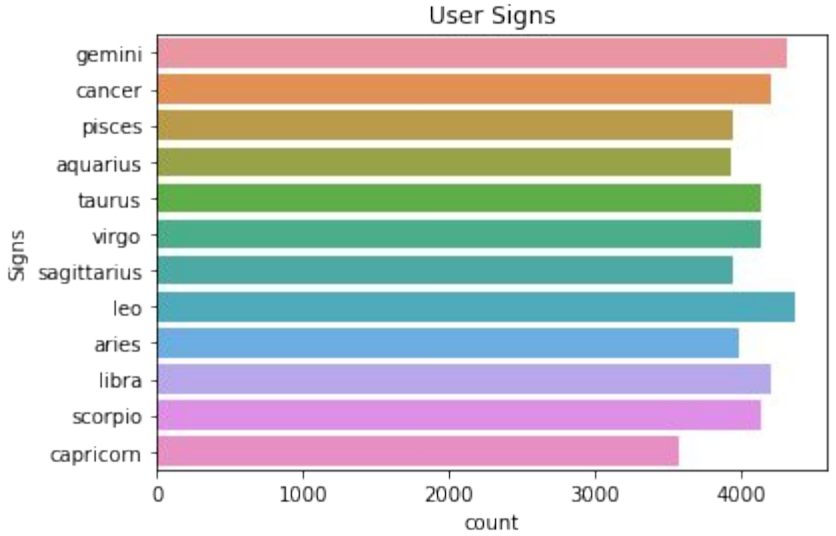


A Look at the Data (cont’d)

- Both ethnicity and religion had needed to have new columns created in order for the data to be meaningful. In both instances I split the data entry after the first space and grouped the entries together.
  - data['religion_cleaned']=data.religion.str.split().str.get(0)
  - data['ethnicity_cleaned']=data.ethnicity.str.split().str.get(0)
- This allowed for qualifiers like “serious” or “not too serious” to be removed creating a meaningful list for purposes of our predictive model.

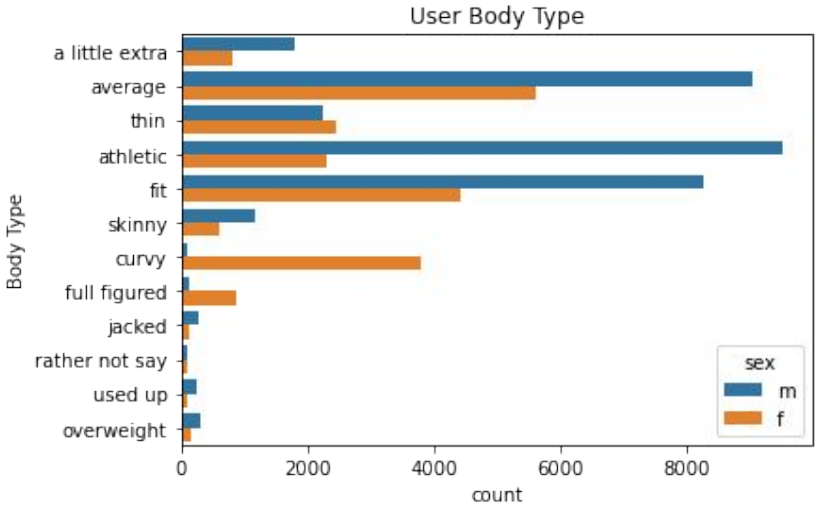
Can an astrological sign be anticipated?

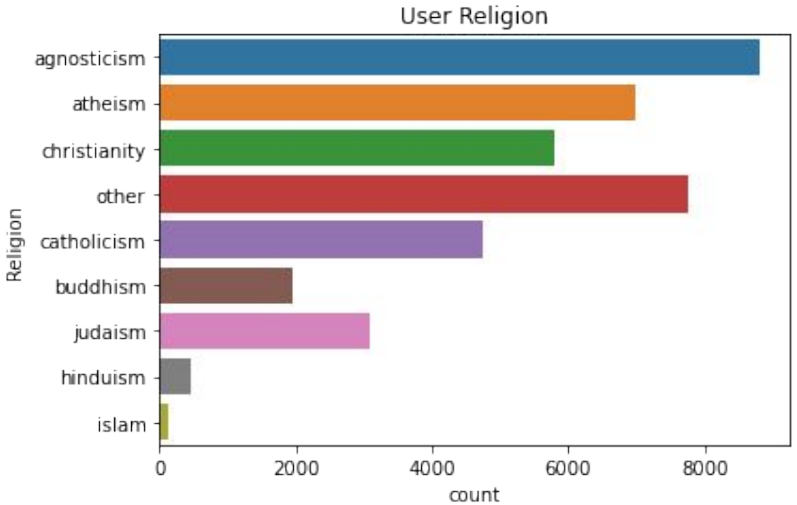
- Because many people take stock in astrological signs, but at the same time many people do not include this data it would be useful to have a method to predict if a person has a certain astrological sign. A model had to be built because birthday was not a field provided.
- In a review of the signs provided by users the data shows an even spread which is particularly important for our purposes.



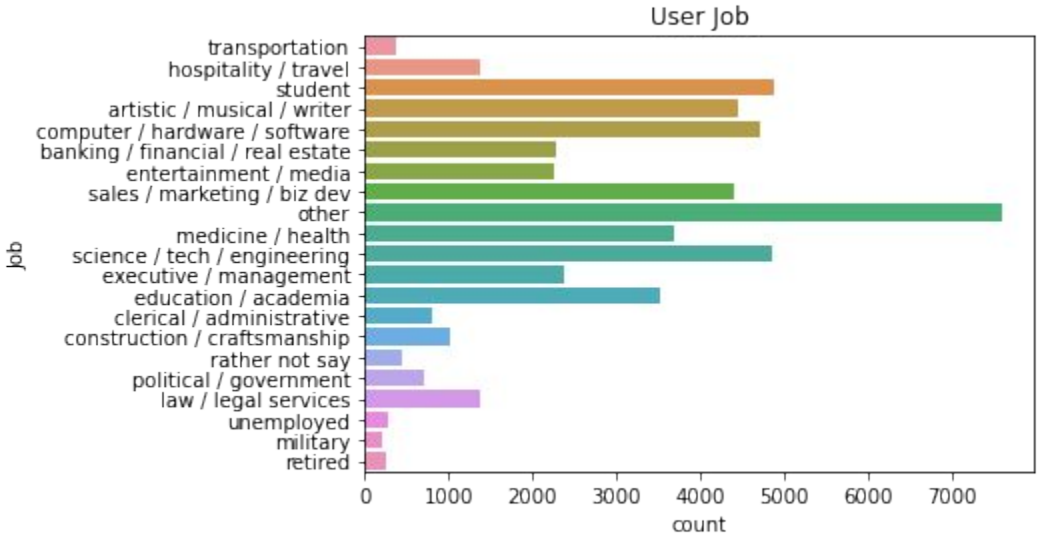
Predictors?

- The following predictors were used either because they were subjective or because they were determined by individual preference.
- These were then added to ethnicity as predictors for our model.





Predictors? (cont'd)



- User language was also used to assist in the process of creating the model. The advantage is that this was so user specific that it gave us a great spread of astrological signs. The downside is that because of the wide variation it is difficult to create a meaningful visualization.

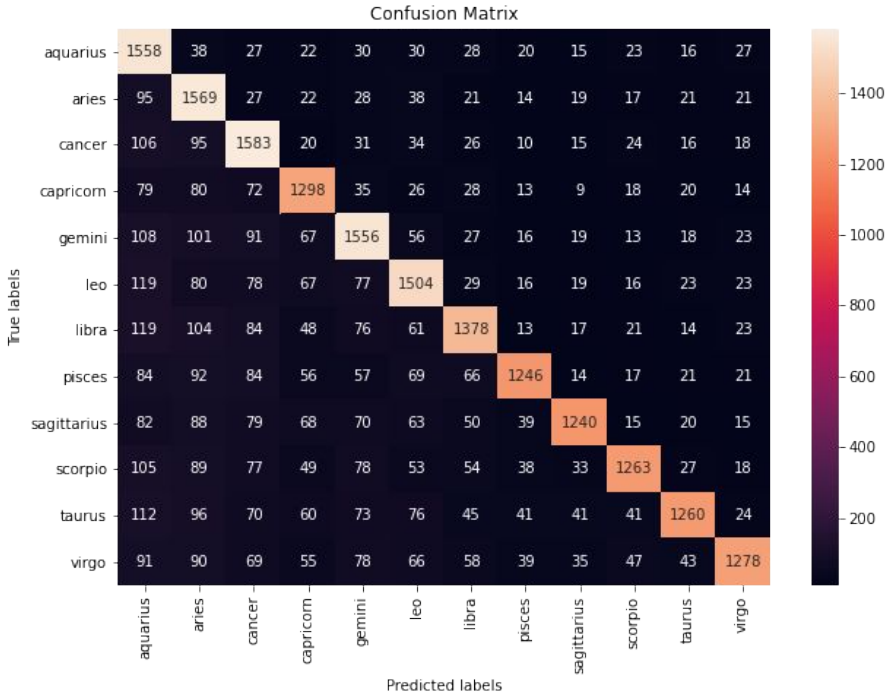
Results

- Using the Decision Tree predictive model and the 5 predictors of ethnicity, job, religion, language and body type we were able to create a model that was able to make predictions of astrological signs with 73% accuracy.

	precision	recall	f1-score	support
aquarius	0.59	0.85	0.69	1834
aries	0.62	0.83	0.71	1892
cancer	0.68	0.80	0.73	1978
capricorn	0.71	0.77	0.74	1692
gemini	0.71	0.74	0.73	2095
leo	0.72	0.73	0.73	2051
libra	0.76	0.70	0.73	1958
pisces	0.83	0.68	0.75	1827
sagittarius	0.84	0.68	0.75	1829
scorpio	0.83	0.67	0.74	1884
taurus	0.84	0.65	0.73	1939
virgo	0.85	0.66	0.74	1949
accuracy			0.73	22928
macro avg	0.75	0.73	0.73	22928
weighted avg	0.75	0.73	0.73	22928

Results (cont'd)

- The confusion matrix also revealed a high rate of success
- K Nearest Neighbor and Logistical Regression models returned results around only 30% (see next slide)



Results (cont'd)

Logistical Regression

	precision	recall	f1-score	support
aquarius	0.27	0.24	0.26	1834
aries	0.26	0.27	0.27	1892
cancer	0.30	0.28	0.29	1978
capricorn	0.47	0.21	0.29	1692
gemini	0.22	0.36	0.27	2095
leo	0.24	0.30	0.27	2051
libra	0.26	0.29	0.27	1958
pisces	0.36	0.22	0.27	1827
sagittarius	0.37	0.24	0.29	1829
scorpio	0.34	0.25	0.29	1884
taurus	0.26	0.30	0.28	1939
virgo	0.23	0.30	0.26	1949
accuracy			0.28	22928
macro avg	0.30	0.27	0.28	22928
weighted avg	0.30	0.28	0.28	22928

K Nearest Neighbor

	precision	recall	f1-score	support
aquarius	0.24	0.62	0.34	1834
aries	0.25	0.49	0.33	1892
cancer	0.29	0.37	0.33	1978
capricorn	0.30	0.33	0.31	1692
gemini	0.33	0.30	0.31	2095
leo	0.37	0.27	0.31	2051
libra	0.37	0.25	0.30	1958
pisces	0.39	0.24	0.30	1827
sagittarius	0.41	0.19	0.26	1829
scorpio	0.38	0.23	0.28	1884
taurus	0.36	0.20	0.26	1939
virgo	0.36	0.21	0.26	1949
accuracy			0.31	22928
macro avg	0.34	0.31	0.30	22928
weighted avg	0.34	0.31	0.30	22928

Conclusion

- While astrological sign is not a perfect indicator of compatibility our model does show that there are ways that a person’s sign can be predicted regardless of whether their birthday is known.
- Ok-Cupid users may find this model useful even if they aren’t able to find the perfect partner on the first try.