Predicting Sex and Income

A study of OK Cupid data

Machine Learning Fundamentals

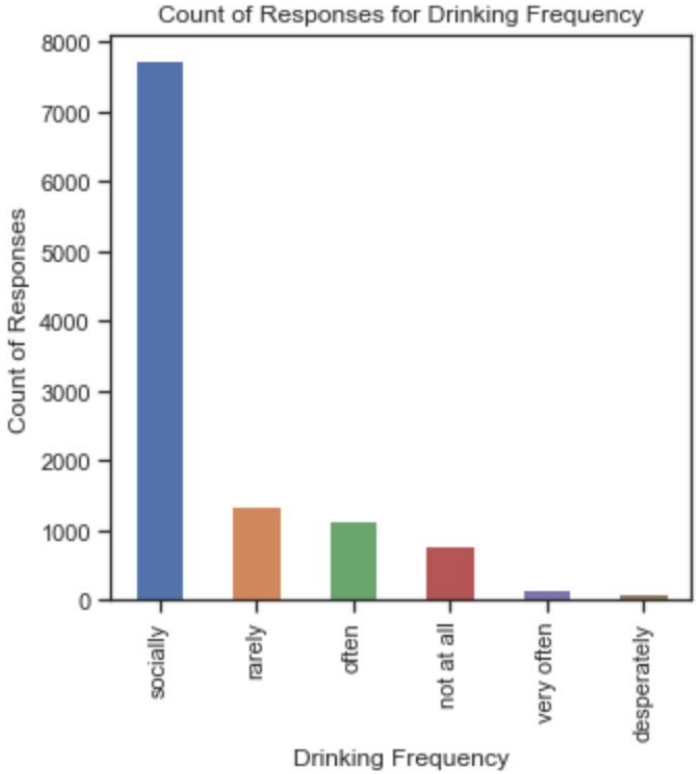
Mike Todd

November 11, 2018

Table of Contents

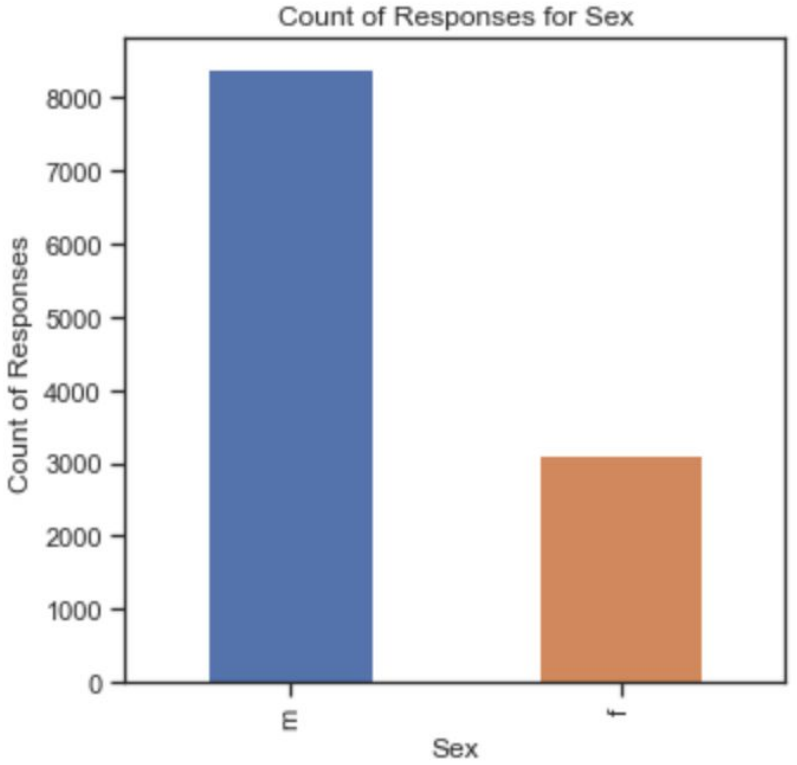
Exploring the data set - Drinks

Response	Count
Socially	41,780
Rarely	5,957
Often	5,164
Not at All	3,267
Very Often	471
Desperately	322



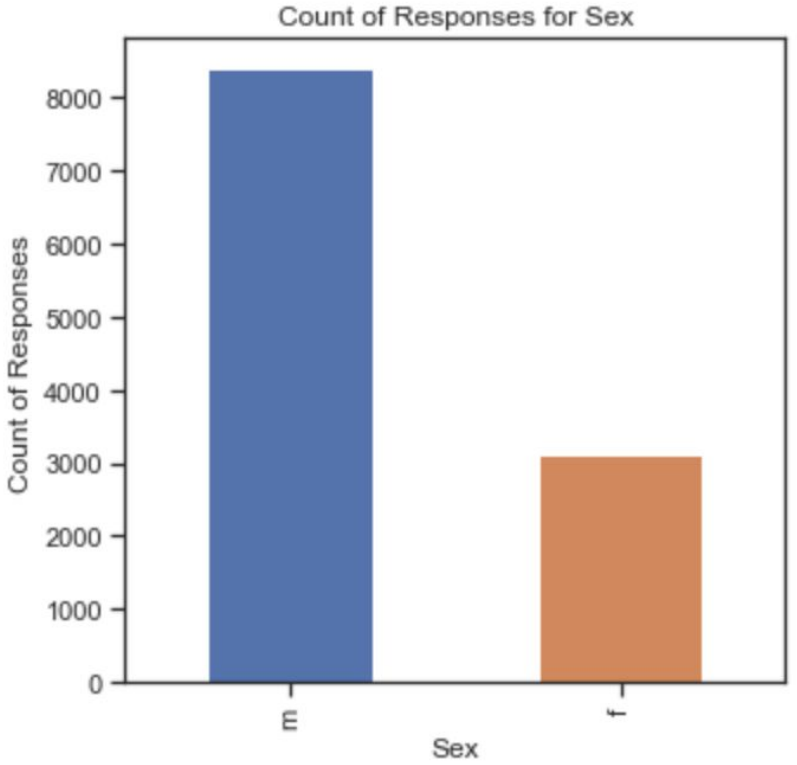
Exploring the data set - Sex

Sex	Count
Female	3,113
Male	8,391



Exploring the data set - Sex

Sex	Count
Female	3,113
Male	8,391



Questions

Can we predict sex using age, height, income, education, and drinks?

Can we predict a person’s income using age, height, sex, education, and drinks?

Augmenting the Dataset - education_years

I mapped the education responses to my best guess of the equivalent education years for each response.

Response	Yr	Response	Yr	Response	Yr
graduated from college/university	8	working on space camp		dropped out of high school	2
graduated from masters program	10	working on law school	9	high school	4
working on college/university	6	two-year college	6	working on high school	2
working on masters program	9	working on med school	10	space camp	
graduated from two-year college	6	dropped out of two-year college	5	ph.d program	12
graduated from high school	4	dropped out of masters program	9	law school	11
graduated from ph.d program	12	masters program	10	dropped out of law school	10
graduated from law school	11	dropped out of ph.d program	11	dropped out of med school	11
working on two-year college	5	college/university	8	graduated from med school	12
dropped out of college/university	6	graduated from space camp		med school	12
working on ph.d program	11	dropped out of space camp			

Augmenting the Dataset - sex_code

I assigned **‘m’** to integer **1** and **‘f’** to integer **0**, so I could include this field in regression analysis.

```
df.loc[:,'sex_code'] = df.sex.map({"m": 1, "f": 0})
```

Classification Approaches - KNN vs Naive Bayes

My classification objective was to determine the sex of respondent based on age, height, income, education_years, and drinks_score.

- **Simplicity** - Naive Bayes was much simpler to run because there was no need to calculate the optimal *k* value, like there was for K Nearest Neighbors.
- **Time to run the model** - Because I created a loop to determine the optimal k value, the model had to run 20 times, so it was slower than the Naive Bayes model.
- **Accuracy/Precision/Recall** - In the end, the K Nearest Neighbors model performed better with an accuracy of 0.828 and a F1 score of 0.886 compared to the Naive Bayes model which earned an accuracy score of 0.718 and an F1 score of 0.836.

Regression Approaches - MLR vs KNNr

My classification objective was to determine the income of the respondent based on age, height, sex, education_years, and drinks_score.

- **Simplicity** - Again, the MLR method was simpler to complete because there wasn't a need for determining the k value.
- **Time to run the model** - Again, because I created a loop to determine the optimal k value, the model had to run 20 times, so it was slower than the MLR model. I also spend more time troubleshooting with this model because it resulted in a negative R-squared value, which surprised me.
- **Accuracy/Precision/Recall** - In this case, the MLR method resulted in a better R-squared value, however it was still extremely low - 0.0046. The best R-squared I could get with the KNN regressor model was -0.3637.

Conclusions

- It seems we **can** predict sex using age, income, education_years, drinks_code, and height with a fairly high level of accuracy.
  - I also found that we can predict sex using just income and education_years with over 70% accuracy.
  - This seems to suggest a pay gap for men and women with the same education background. If I were to explore this further, I'd want to see more data about the industries and positions of the profiles to see how much of this gap is driven by the choice of industry.
  - Also with job title information, I would be interested in comparing the incomes of men and women with the same education and job title.

Conclusions (Continued)

- It seems we **cannot** accurately predict income with age, sex, education_years, drinks_code, and height.
  - I think this is a difficult task for the data provided. The income values appear to be fairly discrete values, and there are many records that do not have valid data for this field.
  - I think these models could be more successful if there were more income responses in the data.
  - Some other data that might help me better answer the question could be zip code data. I expect that people living in the same zip code are likely to have similar incomes.
  - Also if we had data about which profiles were connected (friends), we might be able to use that as an income indicator.