Investigation of the OKCupid Dataset

Utilising Machine Learning

By Michael Kelly

Contents

Question 1: Can we predict the users sex with education level and income?

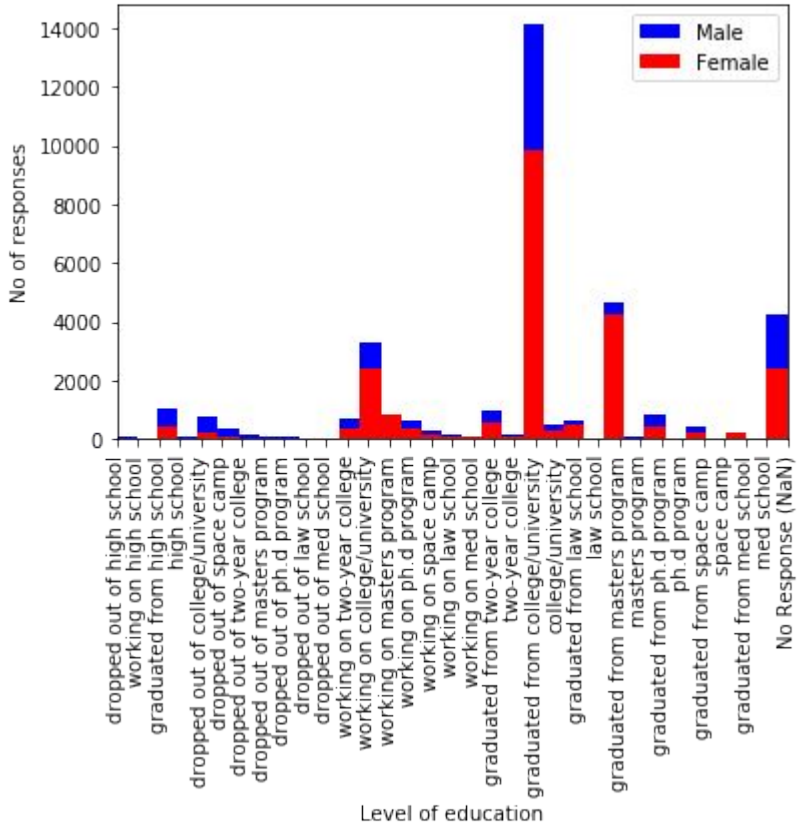
- K-Nearest Neighbor Classifier Approach
  Vs
- Naive Bayes Classifier Approach

Question 2: Can we predict the users age with Income and job?

- Multiple Linear Regression Approach
  Vs
- K-Neighbors Regressor Approach

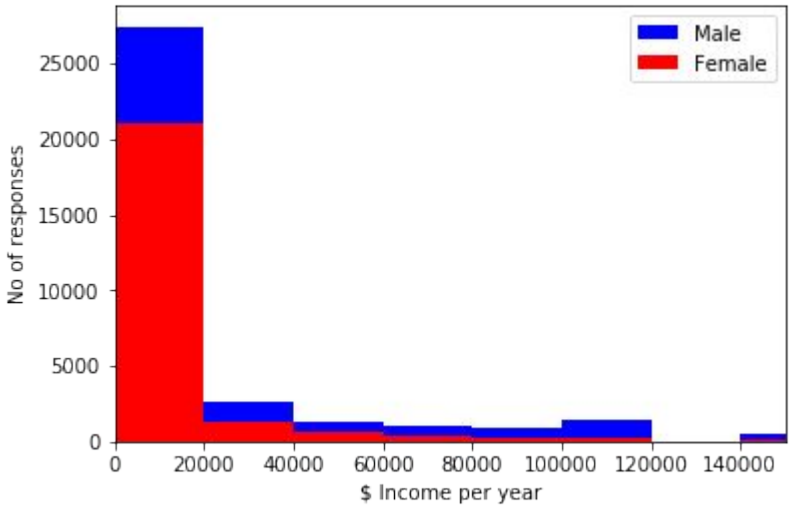
Question 1: Can we predict the users sex with education level and income?

Splitting the dataset columns into male and female components for analysis revealed that while males typically had a higher number of responses (due to the higher number of male users) they followed the same relative pattern to females with regard to education. Though it is notable that relative to number of responses a higher portion of females had graduated from a masters program



Question 1: Can we predict the users sex with education level and income?

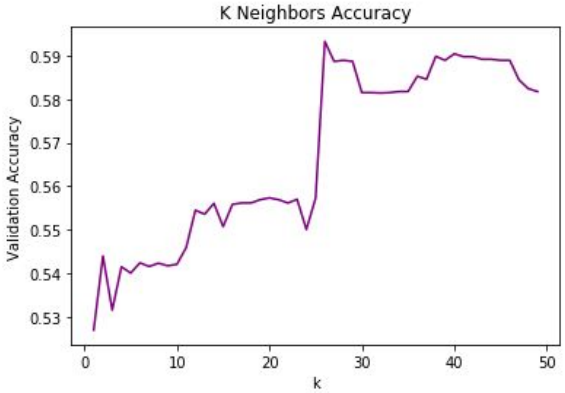
The same can be said for level of income with the vast majority of both sexes opting not to disclose income to the site (indicated as 0 on the graph)



Question 1: Can we predict the users sex with education level and income?

K-Nearest Neighbor Classifier Approach

- Significantly slower to ‘tune’ the right k-value to fit the dataset
- More accurate once tuned




```
Max accuracy found a k=26
```

Accuracy Score: 59.32%

Naive Bayes Classifier Approach

- Simple to create but lacks ability to manually tune
- Slightly less accurate in this case

Accuracy Score: 59%

Question 2: Can we predict the users age with Income and job?

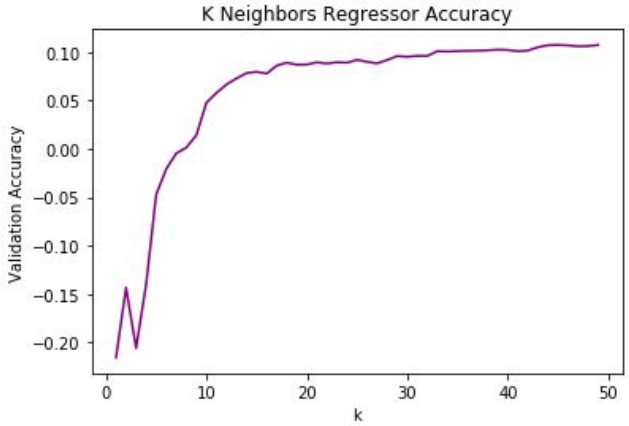
Multiple Linear Regression Approach

- Simple approach and very fast to compute
- Not accurate at all in this dataset

Accuracy Score: 1.66%

K-Neighbors Regressor

- Much slower to compute and the added need to ‘tune’ the k value



Accuracy Score: 8.71%

Closing Comments

Question 1 - the approaches found minor correlations between the inputs (education and income) and sex, which resulted in being able to predict the users sex with only ~60% accuracy (*which by chance is the % of male users). This was also discovered by viewing the data manually, with no discernable difference in responses other than more females graduating with a masters degree as a percentage of users.

Question 2 - the approaches found little to no correlation to be able to accurately predict the users age given the users Income and job with any reasonable accuracy. Though the K-Neighbors Regressor fared significantly better than its counterpart on this dataset, its accuracy was still minimal at ~9%. This could potentially be improved further by exploring a higher number of neighbors.