# Date-a-Scientist Analysis

Codecamey Pro Machine Learning Fundamentals Capstone Project

Zach Goldberg

# Machine Learning techniques were applied to OkCupid profiles to answer questions

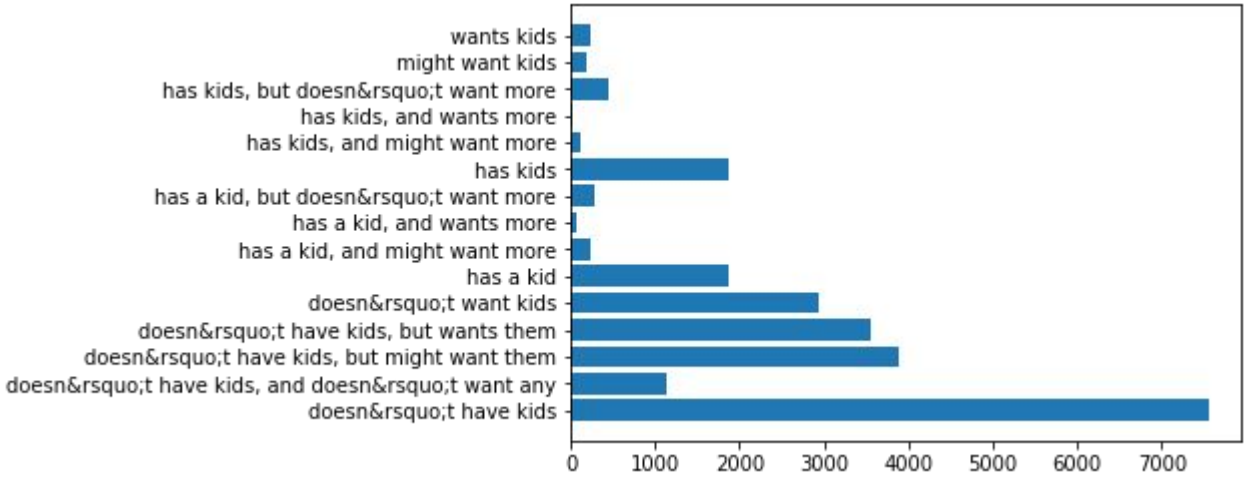
Data was provided from a dating app, OkCupid, which focused on matching users based on multiple-choice and open-ended questions

- Provided data include 18 multiple-choice and 9 essay questions

- Free-from initial exploration of the data drove development of questions to be asked and answered using classification and regression techniques
- Additional techniques were applied to assess the success of the resulting of the resulting answers and models

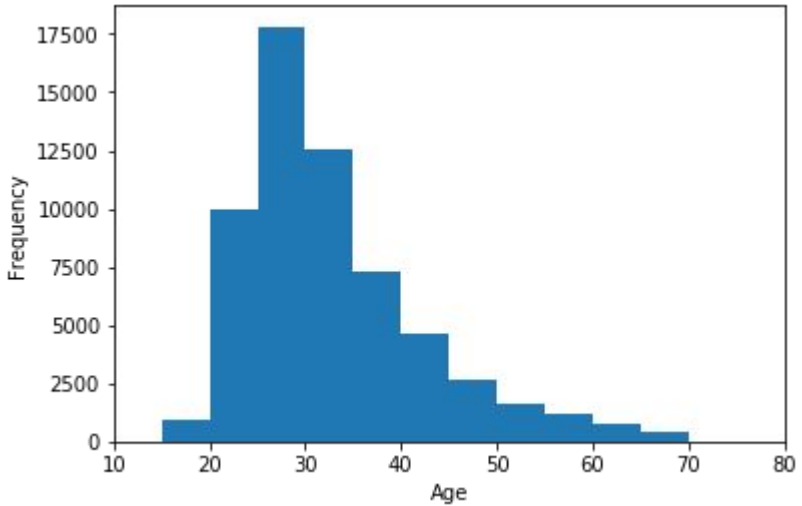
# Data Exploration: Unsurprisingly, not many kids

- Roughly 80% of users do not have children
- Despite the imbalance, I thought it likely that features within the data might help differentiate these two groups
- Creation of a new, binary / aggregated column would be necessary to support this analysis



# Data Exploration: Age range ‘peaks’ interest

Though originally expected to be a much smaller set of ages (in the 20-30 range), the wider (though still left skewed) range of ages indicated that there may be differentiating features in the data that may allow for the development of a regression model to predict age.



# Question 1: Can age, income, and essay length predict whether or not a user has children?

# Question 1: Predicting whether a user has kids

### Why I asked:

- Though only 20% of users had children, I hypothesized that certain features might be be used to predict this scenario
- In the general population, an increase in age increases likelihood of having children
- There is potential that a higher income may correlate with having children (perhaps implying means to care for them?)
- I wondered if a longer essay indicated “more attentiveness / thoughtfulness” and whether this correlated with having kids

### What was needed:

- A new column was needed, summarizing the various options under the “offspring” column, aggregating them to a binary yes/no “has kids?”
- Essays were aggregated and total length was calculated
- All three variables were normalized and split into training and validation sets

# Comparison of classification approaches

## A) K-Nearest Neighbors

**Accuracy Score:** 83.9%

**Speed:** Slower than B

**Simplicity:** Setting up a loop to identify the optimal number of neighbors took a relatively modest amount of effort for significant impact

## B) Support Vector Machine

**Accuracy Score:** 83.8%

**Speed:** Faster than A

**Simplicity:** Plotting the data allowed for quick selection of a linear kernel; however, determining gamma and C was slow and not methodical; unclear if optimal values were selected.

***Assessment: Both models outperformed random guessing and were relatively accurate; despite total run time the precise method of optimizing 'k' makes K Nearest Neighbors a preferred option in this scenario***

# Question 2: Can age be predicted based on drinking/smoking/drug use habits, income, having children, height and essay length?

# Question 2: Predicting age with regression

### Why I asked:

- Age was surprisingly more varied than expected for a dating app
- Age is typically used as an indicator / predictor for other characterstics, so I thought it would be interesting if it itself could be predicted using regression techniques with this dataset

### What was needed:

- Drinking, drug use, and smoking habits were recoded using the methodology shared in the instructions
- Essays were aggregated and total length was calculated
- All variables were normalized and split into  training and validation sets

# Comparison of classification approaches

## A) Multiple Linear Regression

**Accuracy Score:** 31.4%

**Speed:** Negligible

**Simplicity:** Testing the various combinations of features / assessing correlations was somewhat time consuming, with no effect - the model was only weakened by removal of variables.

## B) K-Neighbors Regression

**Accuracy Score:** 29.4%

**Speed:** Negligible

**Simplicity:** Setting up a loop to identify the optimal number of neighbors took a relatively modest amount of effort for significant impact

***Assessment: Neither model was a good predictor of age. Alternate features (use of certain words, perhaps) may fare better, though perhaps best would be bucketing into age ranges and using a classifier instead***

# Conclusions:

1) Income, age, and essay length are decent predictors of whether or not a user has children

2) Selected features were not effective predictors of age

# Next Steps

- Enhance the “has kids” classifier using additional data:
  - Count of the number of time the words “kids” or “children” appear in the essays
- Develop a Naive Bayes classifier to analyze essay text to predict age, rather than previously tested regression approach
- Try approaches to predict what languages a person speaks based on essay texts and location