Machine Learnings

Machine Learning Fundamentals

Benjamin Conrad

January 21, 2019

Disclaimer:

- I did not know python before enrolling in this Intensive course.
- This capstone project proved to be significantly more difficult for me than the rest of the course.
- It took me an entire week to explore the data and make sure I was working with it correctly *

* I'm still not sure I am working with the data correctly

Table of Contents

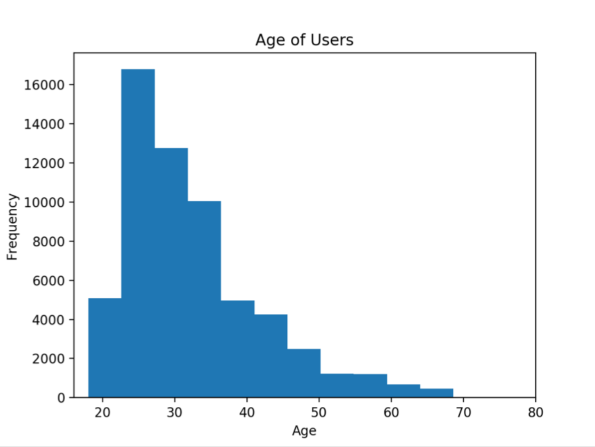
- exploring the data
- questions and concerns
- data augmentation: wins and failures
- attempts at classification
- attempts at regression
- conclusions and wishes

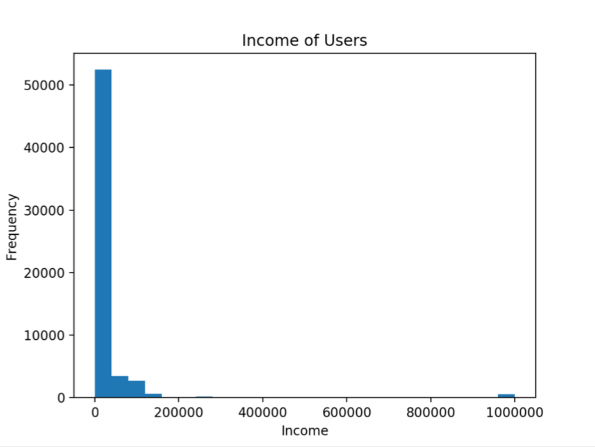
Exploring the Data

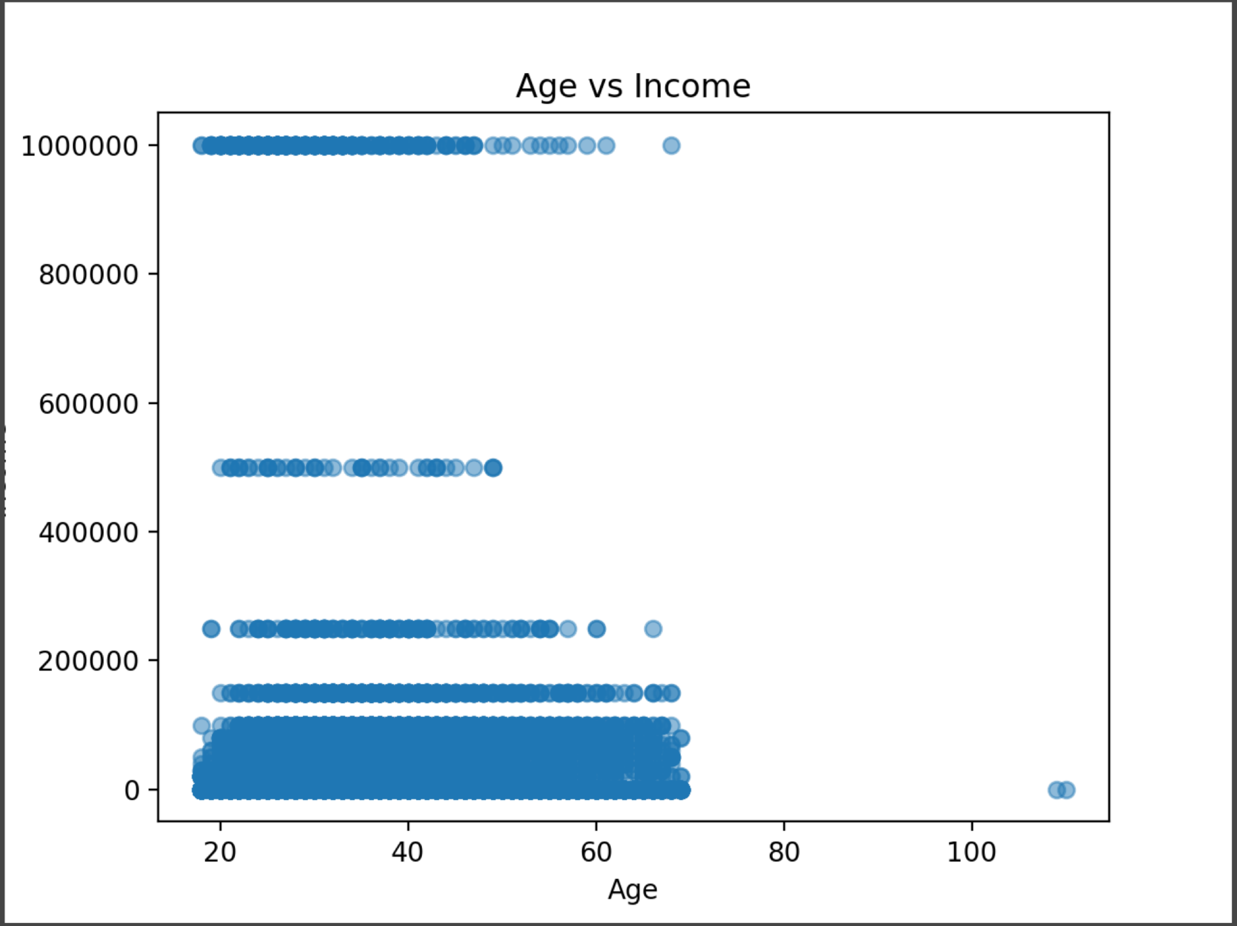
Value Counts!

Upon creating a dataframe from the data in profiles.csv, my first step in data exploration was to print the `value_counts()` of each and every column. I was initially curious about the columns regarding zodiac sign as well as offspring. What was interesting to me about this data was that they didn't only ask what a users sign was, but how important that was to them. Similarly, the offspring column provided a wealth of information that I was excited to make graphable. However, as neither of these columns had numerical data, I first focused on plotting users age, income, and the two against each other. As you will see in the next slide, the results are underwhelming.

Age vs Income







Questions & Concerns

Who's making babies?

After taking a look at all the data that was available to me, I came up with a number of questions. Inspired by the movie 'Idiocracy' I wanted to try to answer the following questions:

- Can we predict whether a user has children based on age, level of education, income, and other factors?
- Can we predict whether a user wants, but does not have, children based on similar factors?
- Can level of education be predicted based on the Flesch-Kincaid Grade Level of the users essay answers?
- Are users with higher levels of education more or less likely to have children than users with lower education levels?

Data Augmentation: Wins & Failures

Data Wins

I created many new columns to help me try to answer these questions. I first started by adding the same columns described in the instructions, as I thought it couldn’t hurt to use the drinks, smokes, and drugs columns in my models. I also created columns to describe a users educational level, whether they have children, whether they way children, whether they have pets, and how important their sign is to them regardless of what their sign is.

Got Kids?

Below is a table describing how I made the 'has_kids_code' column. I first had options going from 0-3 for 'not specified', 'has no kids', 'has a kid', 'has kids'. However, in practice this did not yield great results, so I changed it to a simple 1 for 'has at least one kid' or a 0 for 'has no kids or did not specify'. I did the same thing to create the column 'wants_kids_code'.

User	offspring	has_kids_code
0	has kids, and wants more	1
1	might want kids	0
2	has a kid, but doesn't want more	1

Data Failures

As previously mentioned, I hoped to create a column representing the Flesch-Kincaid Grade Level of each users combined essay answers. I had found a library called py-readability-metrics that seemed like it could do the job, but after much trial and more error I had to abandon it. In the end, the run time was much too long and the data produced did not appear to be accurate. I have left the code commented out in my file for further review.

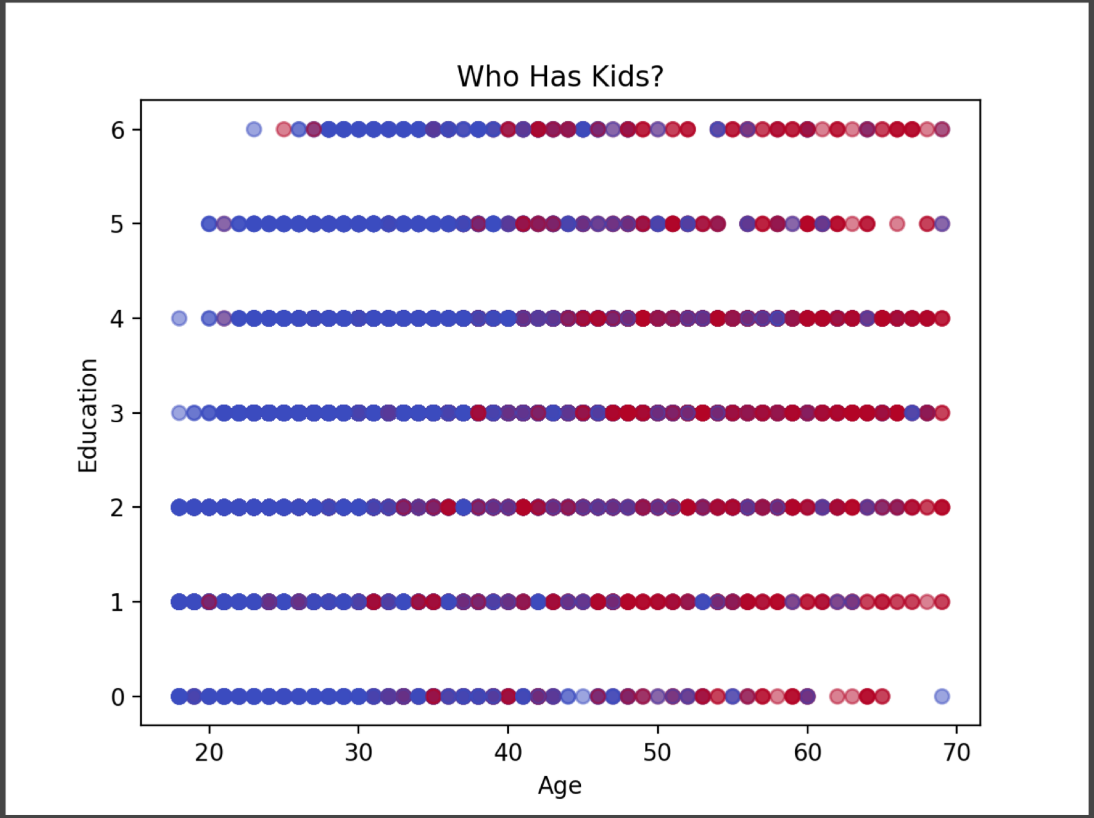
Attempts at Classification

Classification with Support Vector Machines

I first tried to predict whether a user had a child using their age and level of education. This graph represents users with kids in red and users without in blue. I tried many methods of plotting the decision boundary of my model but was unsuccessful. Despite my best efforts, I was also unsuccessful at getting sklearn.metrics.classification_report to work. I played round with various parameters and the best scores I was able to get were:

Train Score: 0.8581

Test Score: 0.8351

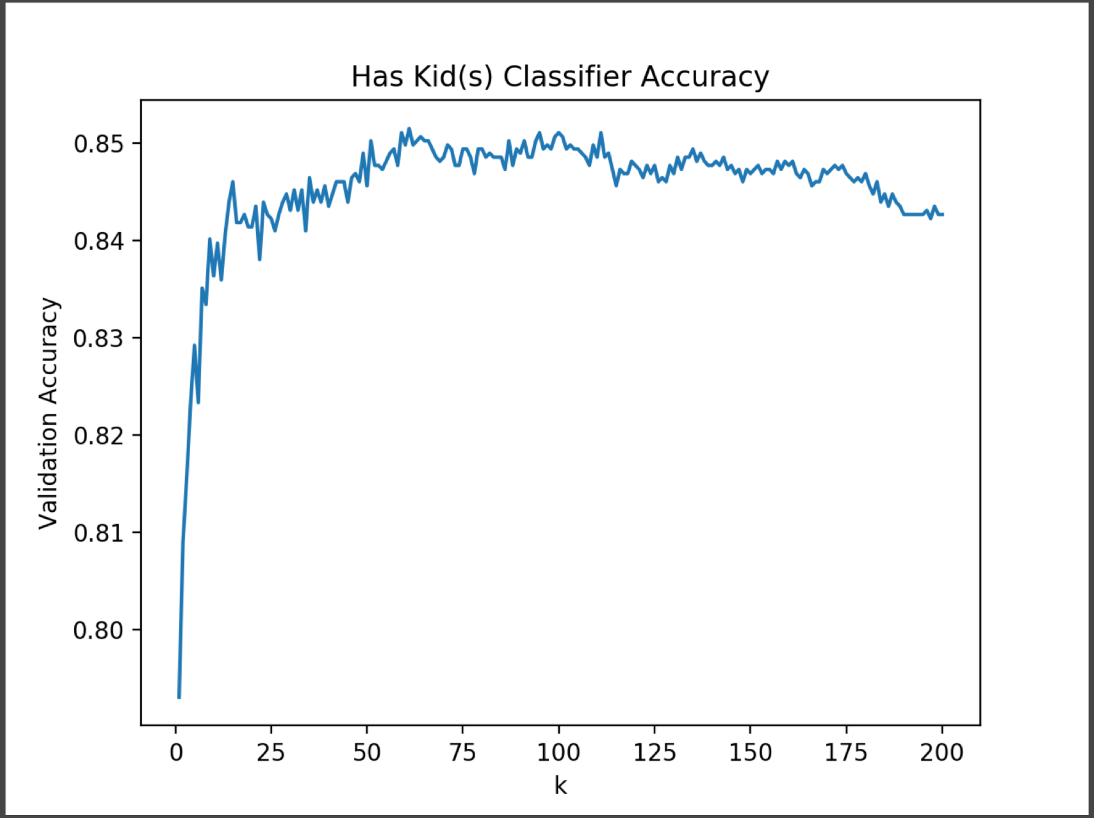


Classification with K-Nearest Neighbors

To approach the question form a different angle, I attempted classification with k-nearest neighbors. This time, instead of making a prediction using only age and education level, I used all the features from my feature_data dataframe (see code for full list). While the run time took longer than the svm approach, the accuracy was not equally increased.

Train Score: 0.8586

Test Score: 0.8515



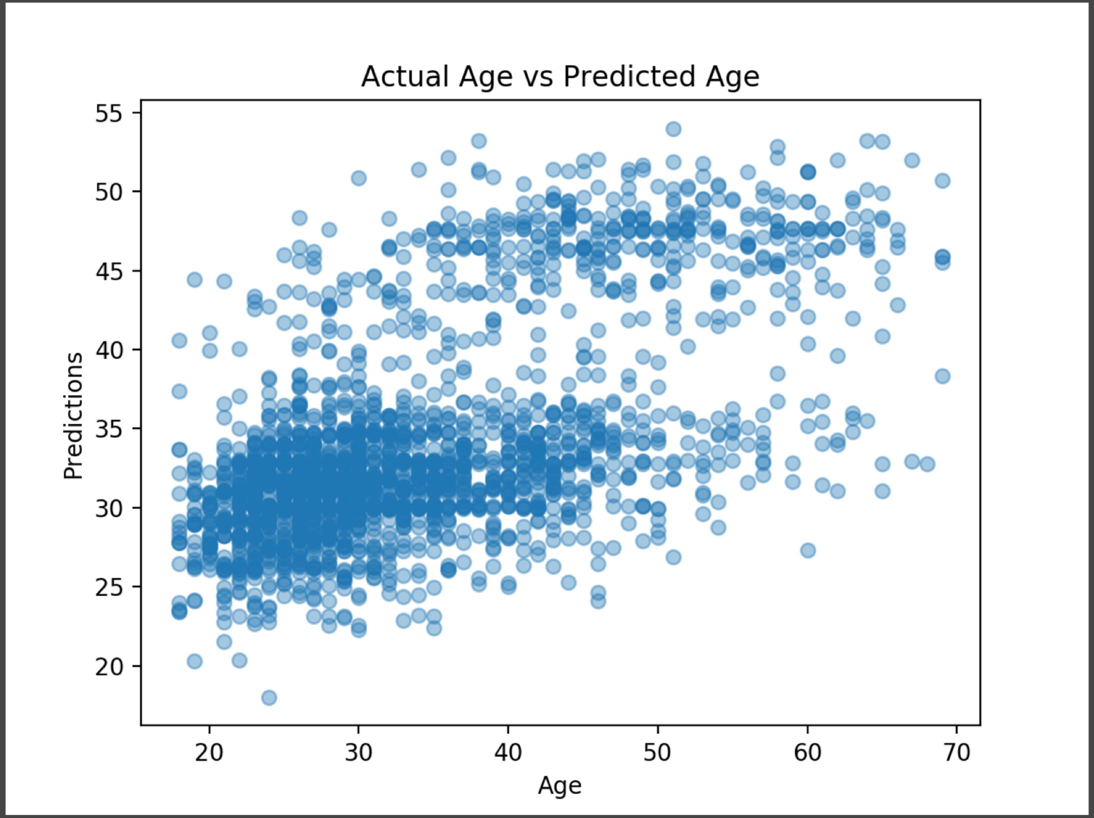
Attempts at Regression

Regression with Multiple Linear Regression

Similar to my attempts with k-nearest neighbors classification, I again used all features to make a prediction. I tried to predict many different things from who had kids, to who wanted kids, to sign importance. I reviewed coefficients and removed and adjusted features included. No matter what, I could not get my accuracy above the scores below, which predict age based on all features.

Train Score: 0.3790

Test Score: 0.3645

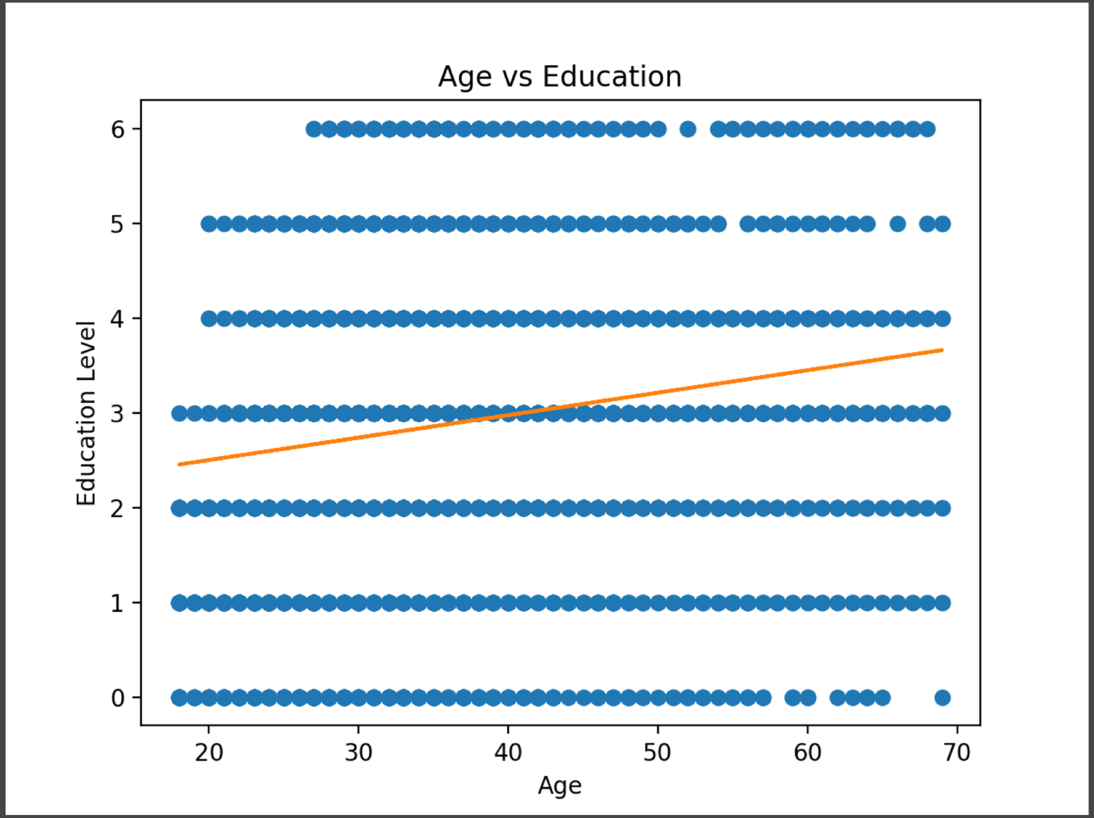


Regression with Linear Regression

After my attempts with multiple linear regression, I tried regular linear regression and k-nearest regression. Both had unimpressive results. Again, I could not seem to find the right question to ask that linear regression could help to answer. The graph on the right and scores below represent the abysmal performance of predicting a trend between age and education level using linear regression.

Train Score: 0.0510

Test Score: 0.0505



Conclusions & Wishes

Conclusions

After crunching the numbers and running the data through numerous machine learning models, here are some rough conclusions and answers to the questions I initially asked:

- **We can we predict whether a user has children based on age.** All other features did not offer additional insight.
- **We can we predict whether a user wants, but does not have, children based on similar factors.** While not saved in my file, the results were equally accurate to the first question.
- **I was unable to make predictions based on the Flesch-Kincaid Grade Level of the users essay answers.**
- **Users level of education had no bearing on their likelihood of having children.** We appear to be safe from Idiocracy!

Next step,
more detailed data!

In my opinion, here are some of the next steps I could imagine taking to better answer these question:

- Collecting data from other sources, not just users of an online dating platform.
  - Perhaps census data or something similar.
- Less multiple choice data.
  - I enjoyed using the age column of our data as it was most varied.
  - Opposingly, I wish the income column could have been as diverse and robust.
  - Getting a number of users children rather then 'either one child' or 'more than one child' would have been interesting to play with.

Thanks for the knowledge!

Machine Learning Fundamentals
Benjamin Conrad
January 21, 2019