Capstone Project: Date a Scientist

Machine Learning Fundamentals

William Rogers

1/21/19

Table of Contents

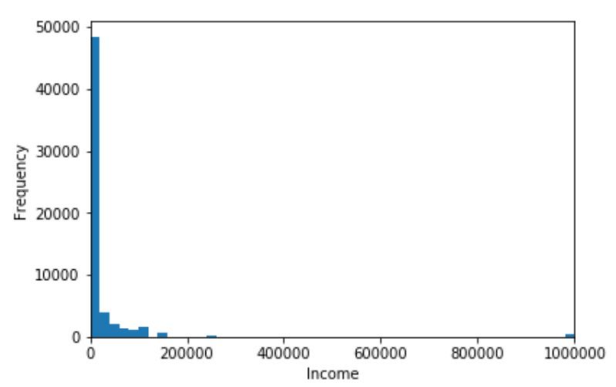
- Questions to Answer
- Exploration of the Dataset
- Augmenting the Dataset
- Classification Approaches
- Regression Approaches
- Conclusions/Next Steps

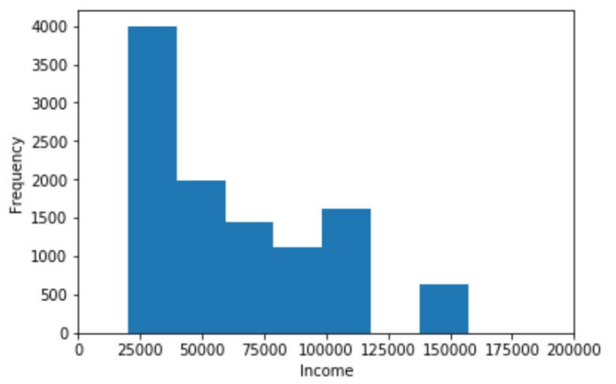
Questions to Answer

I used two questions from the assignment.

- Classification
  - Can you predict a person’s sex using their education level and their income?
- Regression
  - Can you predict a person’s age based on the number of “me” or “I” in their essays?

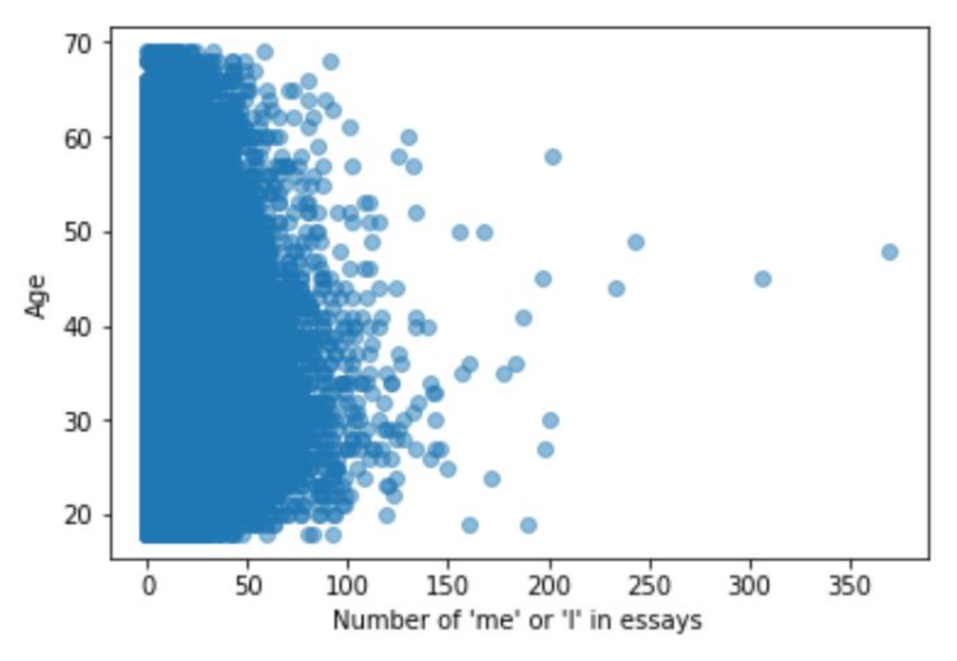
Exploration of the Dataset





I was interested in the income distribution. First I removed all the people who didn't enter an income (-1) andI tried it with 10 bins, but realized the people who put 1 million threw it off. The result was the histogram in the upper left. I then increased the bins to 50, and added an x-axis maximum of with a cutoff of $200k, which resulted in the plot on the left.

Exploration of the Dataset (cont.)



To further investigate my regression question, after I created a new column with the number of “me” and “I”s in the essays (discussed in the next section) I decided to plot these against their age.

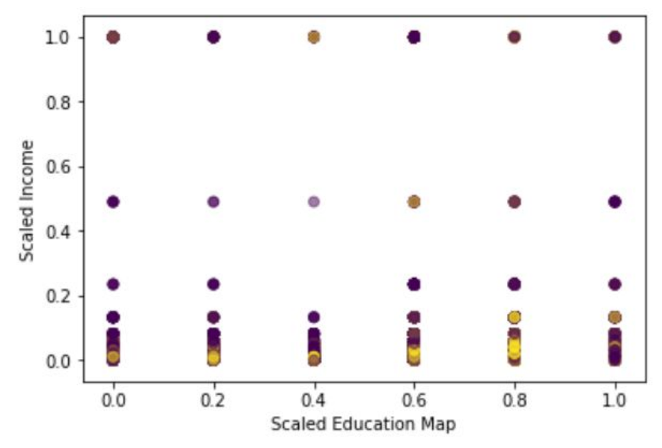
Augmenting the dataset

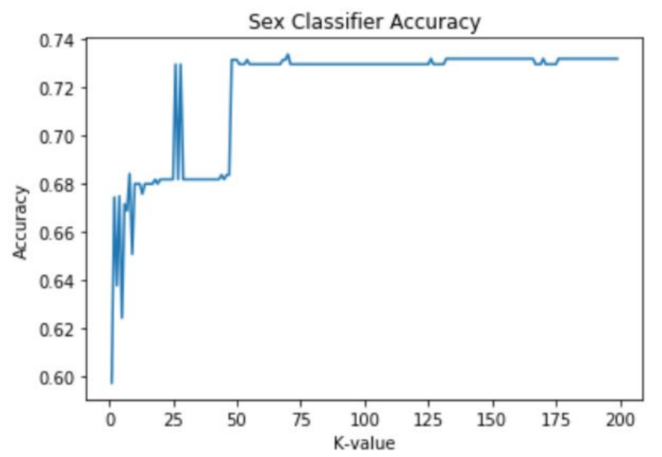
- As mentioned on the last slide, one of the new columns I added was a count of the number of “me” and “I” a person used in all of their essay questions.
- To answer my classification question, I needed to create two more columns. The first was I needed to map the “sex” column, which had either an “m” or “f” for each person, to numbers. I used “m”=0 and “f”=1.
- I also had to make a new column for the education mapping. There were a lot of options a user could select for their education level. I decided that anything below high school would be a base of 0 (this included space camp related answers). Next I attempted to classify them in an increasing difficulty. Graduating high school was a 1, two year college a 2, college/university a 3, masters program and law school a 4, phd and med school a 5. Next I decided that if the answer included “dropped out of”, I would round down to the most appropriate response (i.e. dropped out of two-year college equaled graduated high school, or dropped out of phd equaled finished Masters degree), however if the answer was “Working on <level>” then I rounded up and included it with the graduated category. This resulted in my “education_mapping” in my code.

Classification Approaches

Question: Can you predict sex from a person's education level and income?

- Before running any models, I removed all of the data sets that had no income value. I also normalized the data, and I split the resulting set into training and test sets. A scatter plot of the data, with colors indicating sex can be seen on the upper right.
- K-Neighbor Classifier
  - I ran K-Neighbor classifiers with values of k from 1 to 200 and fit the model to the training sets, then I recorded the accuracy score the model received on the test sets, as well as the time it took to run. The resulting plot is shown on the lower right.





Classification Approaches (cont.)

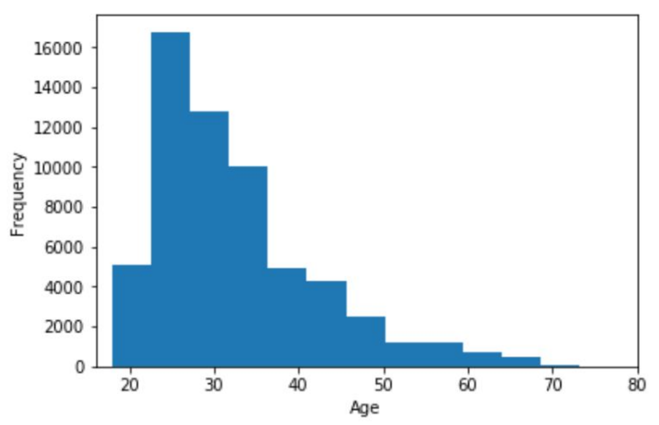
- K-Neighbor Classifier (cont.)
    - **The highest accuracy resulted from a K=69, and was 73.34%. It took 0.1 seconds to run the optimal k-value model, however it took 17.61 seconds to run all 200 models.**
    -
- Support Vector Classifier
    - I also created a SVC model using the same training and test sets. The SVC had an accuracy of 72.69%, and took 4.412 seconds to complete.

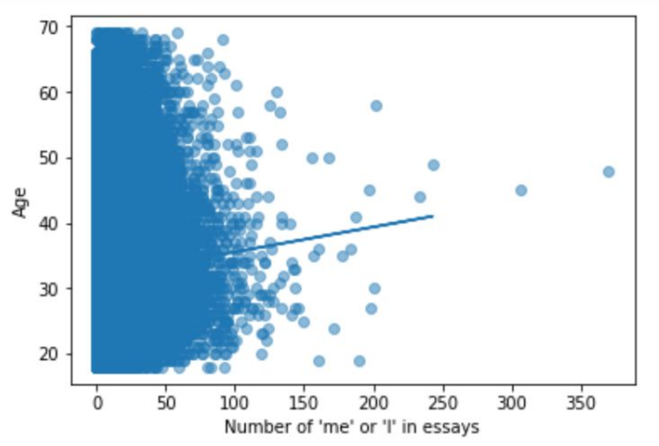
Both the SVC and the K-Neighbor classifier had very similar accuracy values. The single optimal K-Neighbor model was much faster than the SVC, however the running all 200 K-Neighbor models took approximately 4 times as long as the SVC.

Regression Approaches

Question: Can you predict a person's age from the number of "me" and "I"s used in their essays?

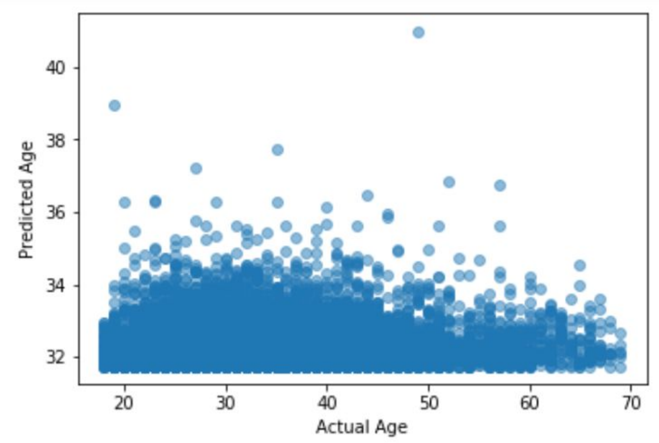
- The only changes I made to the data for this analysis was to remove the people with ages listed over 100. I did not scale the data, as there was only one independent variable.
- Linear Regression
  - The first regression model I ran was a linear regression. The model did not have a very good R^2 value. It's score on the training set was 0.0036, and it's score on the test set was 0.004836. It took 0.088 seconds to run. The resulting line can be seen plotted on the scatter plot to the right.

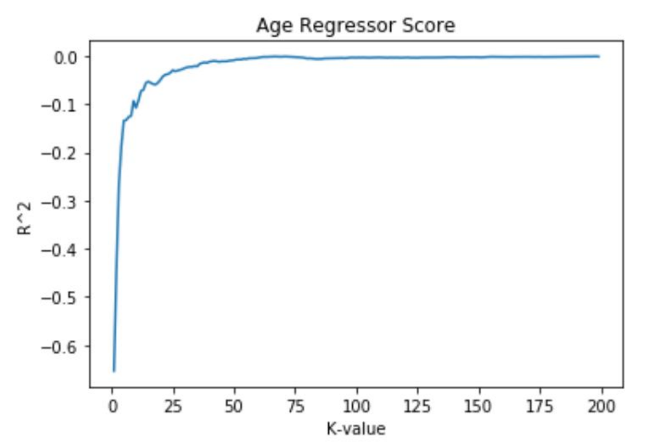




Regression Approaches (cont.)

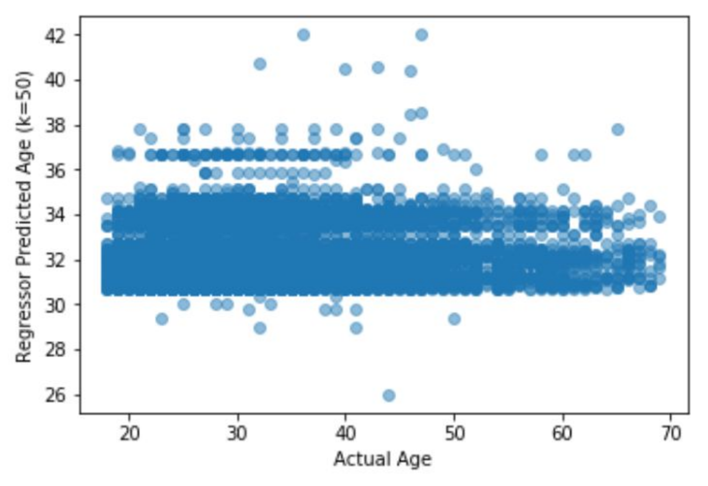
- Linear Regression (cont.)
  - A plot of actual vs. predicted age can be seen on the upper right.
- K-Neighbors Regressor
  - Next I ran a K-Neighbors regressor model for k values of 1 through 200 on the training set, and graphed the calculated R^2 score on the test set. As you can see in the chart to the right, the models were not very good, and resulted in negative R^2 values. In addition to this, these models too much longer to run, with a min and max time of 0.3845 and 0.57 seconds, and totaling 78.429 seconds for all 200.





Regression Approaches (cont.)

- K-Neighbors Regressor (cont.)
  - As there was no especially good K-value for this model, I chose K=50 to graph the actual vs predicted ages, shown on the right.



Neither of these models were particularly good at predicting a person's age, however the linear regression model was much faster than the K-Neighbors regressors.

Conclusions

Classification Question: Based on the results, it would appear that you can guess a person's sex based on their income and education level. Both models resulted in an accuracy of approximately 73%, which is much better than the 50% chance you'd have of guessing a person's sex without any other information.

Regression Question: Based on the results of these models, it does not appear that there is any correlation between a person's age and the number of times they use the words "me" or "I" in their essay responses. As can be seen from the plots, both models do not predict very many ages outside of the range of 30-38 years old. This makes sense, as when looking at the histogram of ages, the vast majority appear to be in the 25-40 years old range.