Date-A-Scientist Project

Machine Learning Fundamentals

Joe Bieniek

Question

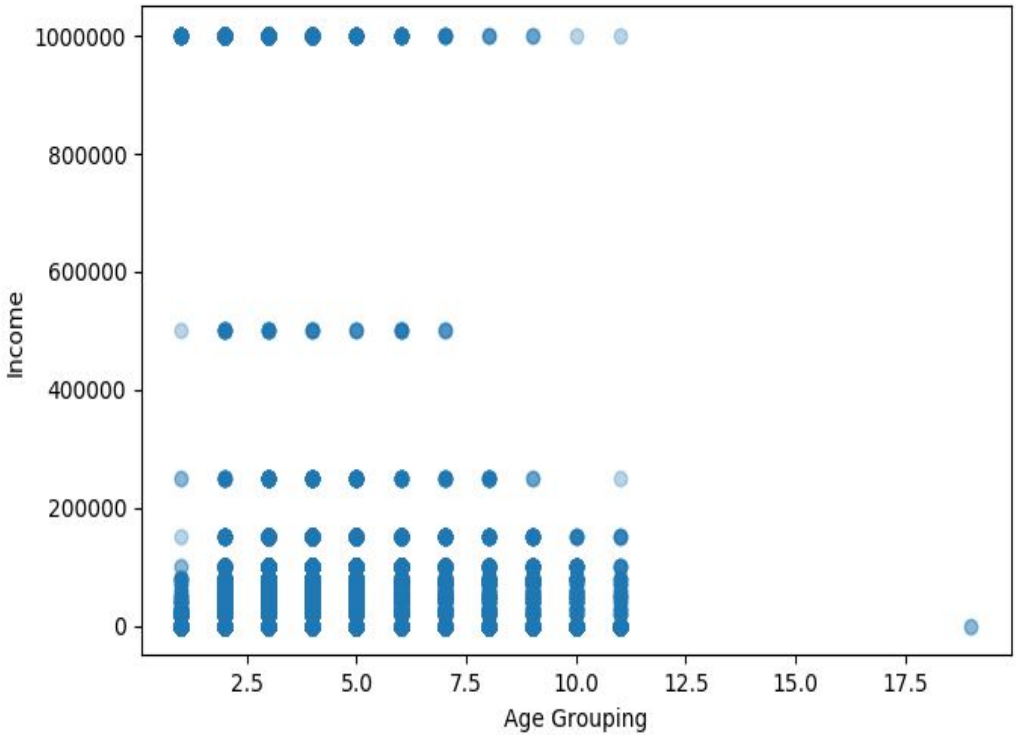
I wondered if you could predict age by looking features that are commonly perceived as turn offs i.e. drinking, smoking, drugs, body type. I was also curious to include income to see if money helps overcome the other deficiencies. My psuedo-hypothesis was that there may be predictability that older singles may have higher density of “negative” features and modest incomes to remain in the dating pool.

Data Modification

I wanted to look at responses to drinking, smoking, drugs, body type, income and age and did the following data transformations.

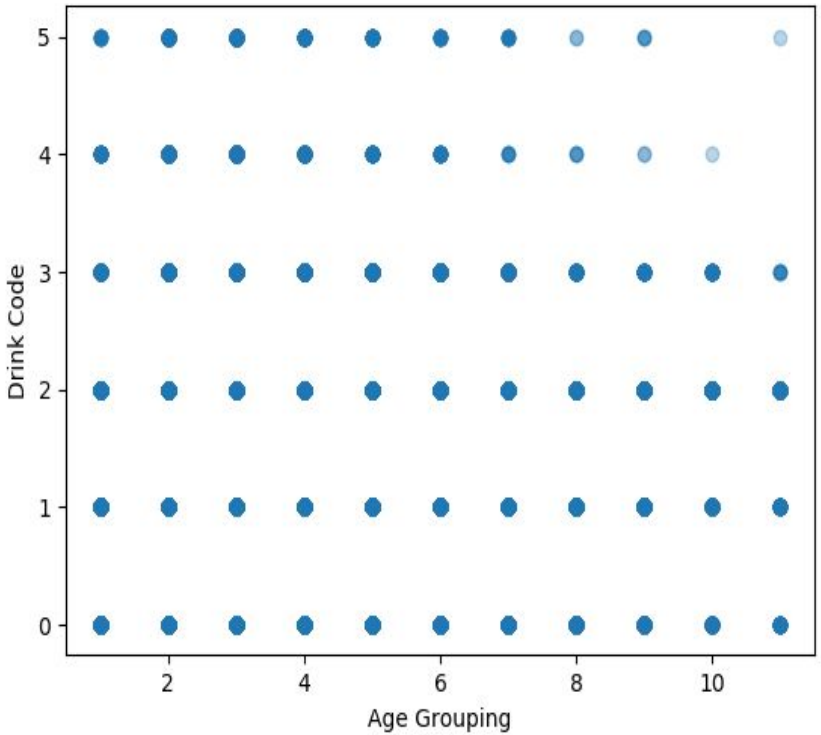
1. Converted age into age groups 20-25, 26-30, etc.
2. Map Drinks values to numeric values
3. Map Smokes values to yes/no. I decided to treat any none “no” answer as “yes” because I think smoking is very polarizing.
4. Map Body Type values to numeric values
5. Map Drugs values to numeric values

Data Exploration



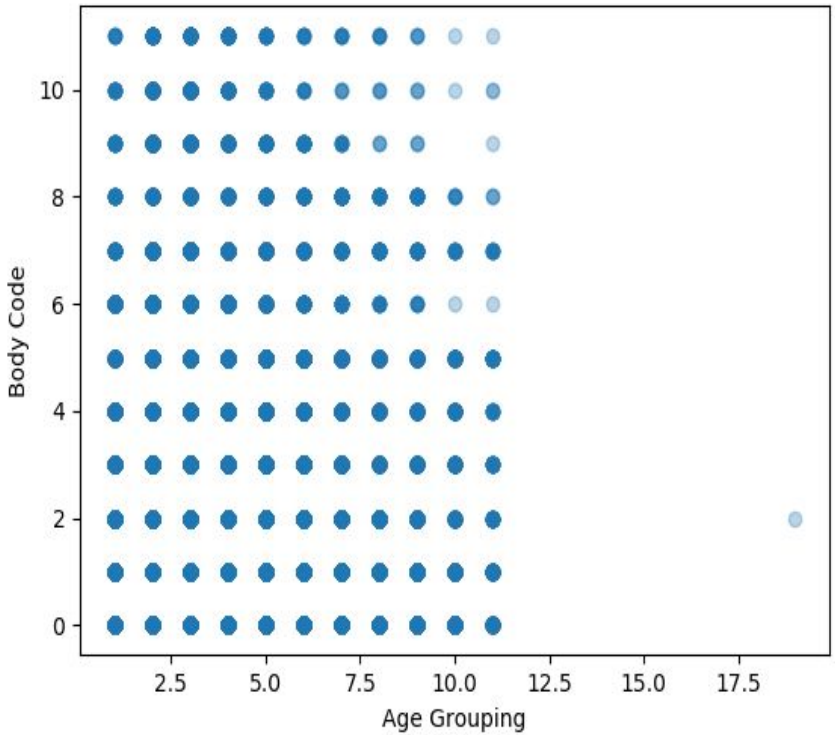
Graphed income by age group. I grouped ages into 5 year ranges

Data Exploration



Graphed drink codes by age group. I grouped ages into 5 year ranges

Data Exploration



Graphed body type code by age group. I grouped ages into 5 year ranges

Data Exploration

After exploring the data some, there didn’t seem to be any real strong correlations to any individual feature I was looking. That left me a bit dubious about my question but thought I’d see what came of it.

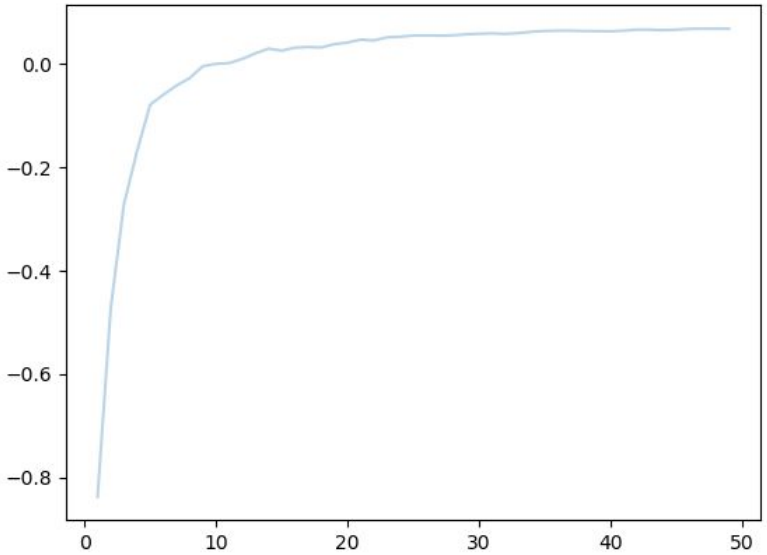
Regression - Linear

I ran a multi feature linear regression and got a very poor accuracy score of 0.008240019136679044

It was simple and quick to run the linear regression, however, the results are totally unusable.

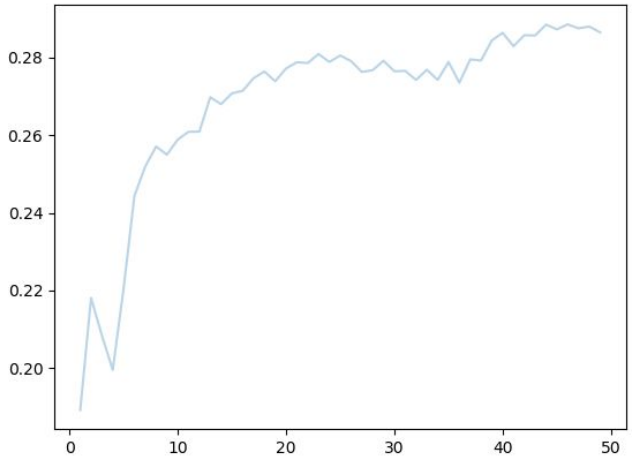
Regression - K Nearest Neighbors

I ran a KNN regressor over 1-50 K and charted the result below. The results were also terrible.



Classification - K Nearest Neighbors

I ran the KNN classification for 50 iterations with the resulting graph below. It maxed out around K=44 for accuracy of 0.2884230427046263. Which still is not very good, but at least getting some results.



Classification - SVC

I also ran a SVC classification. I only attempted a few gamma/C combinations because it took a few minutes to run. It did produce the best accuracy result with 0.2919261565836299, but the marginal increase in accuracy probably doesn’t warrant the extra compute time.

Conclusion

In the end my experiment did not show that there was much predictability for age group with my selected features. I think in the end a lot of the features are too broadly distributed across all age groups to have any real predictive abilities. I think I would have to dig into some of the essays to try and extracts activities and interests to potentially get some better results for age group classification.