Date-A-Scientist Outcomes & Correlations

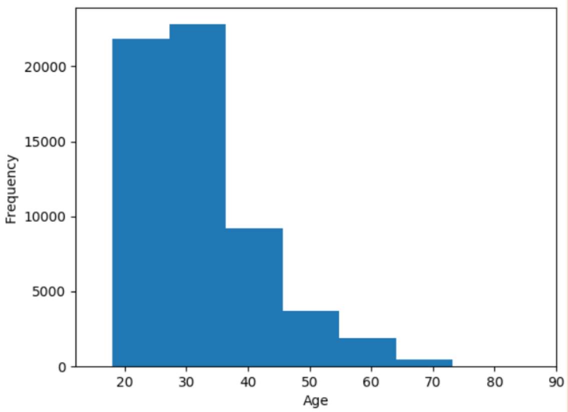
Machine Learning Fundamentals
Ray Anderson
02/2019

codecademy

Table of Contents

- Exploration of the Dataset
- Question(s) to Answer
- Augmenting the Dataset
- ***Responding to Questions***: Classification Approaches & Regression Approaches
- Conclusions/Next steps

Exploration of Data

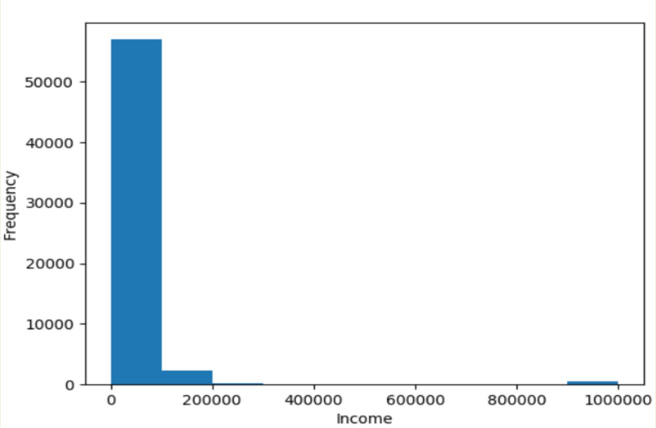


a.

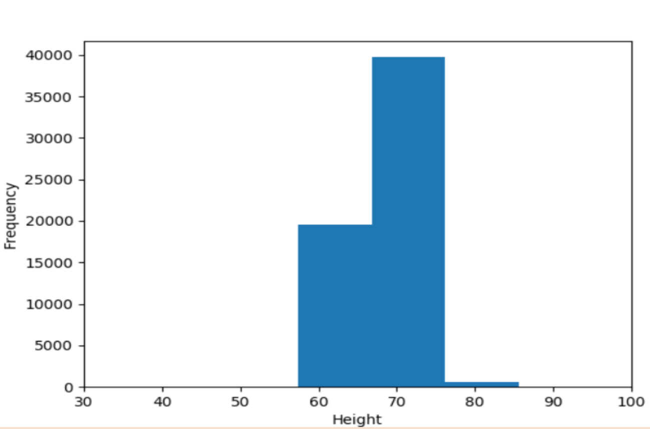
a. Frequency of Users vs. Age

b. Frequency of Users vs. Income

c. Frequency of Users vs. Height



b.

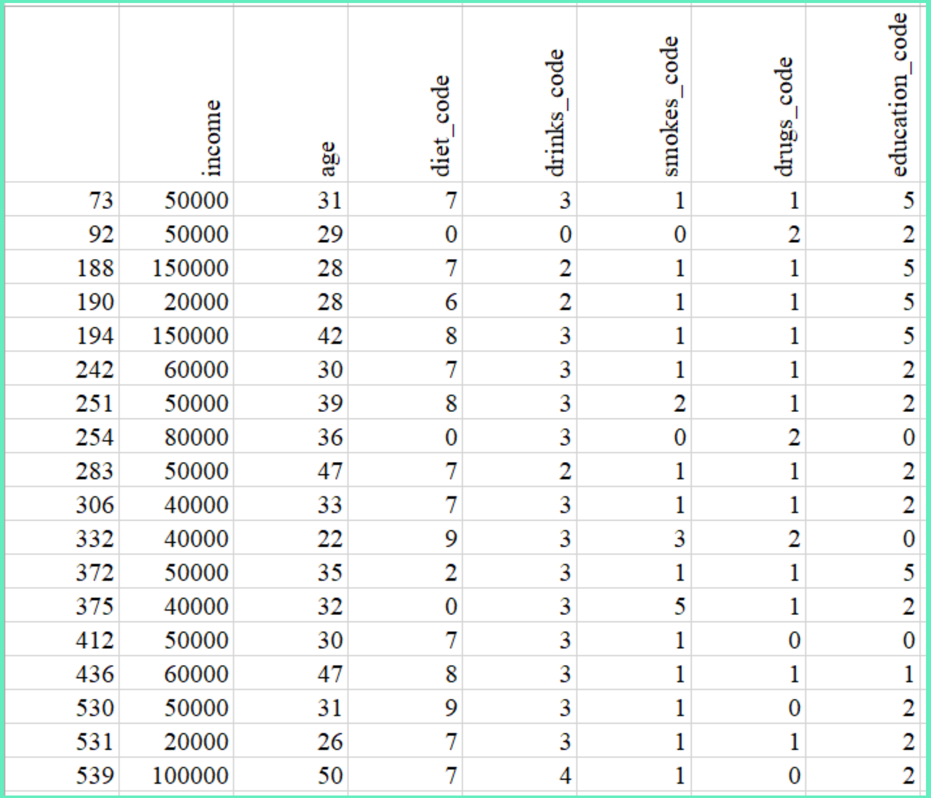


c.

Question(s) to Answer

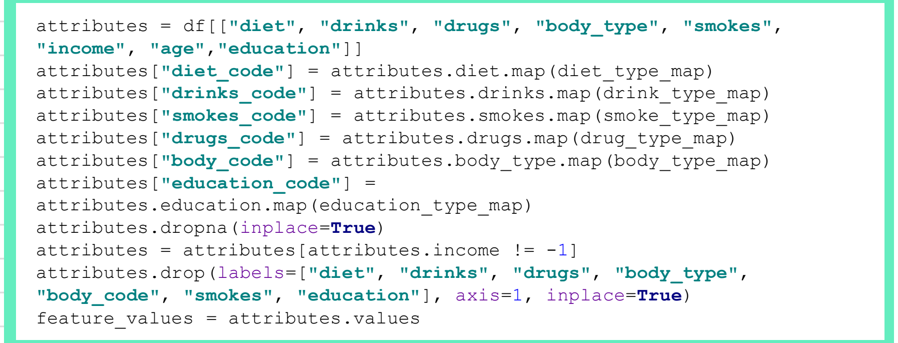
1. What dependency does height have on income? ***(Linear Regression)***
2. What dependency does age have on income? ***(Linear Regression)***
3. What dependency does height and age have on income? ***(Multiple Linear Regression)***
4. What dependency does the frequency of smoking and drinking have on someone’s education level? ***(Naïve Bayes Classification)***

Augmenting the Dataset



	income	age	diet_code	drinks_code	smokes_code	drugs_code	education_code
73	50000	31	7	3	1	1	5
92	50000	29	0	0	0	2	2
188	150000	28	7	2	1	1	5
190	20000	28	6	2	1	1	5
194	150000	42	8	3	1	1	5
242	60000	30	7	3	1	1	2
251	50000	39	8	3	2	1	2
254	80000	36	0	3	0	2	0
283	50000	47	7	2	1	1	2
306	40000	33	7	3	1	1	2
332	40000	22	9	3	3	2	0
372	50000	35	2	3	1	1	5
375	40000	32	0	3	5	1	2
412	50000	30	7	3	1	0	0
436	60000	47	8	3	1	1	1
530	50000	31	9	3	1	0	2
531	20000	26	7	3	1	1	2
539	100000	50	7	4	1	0	2

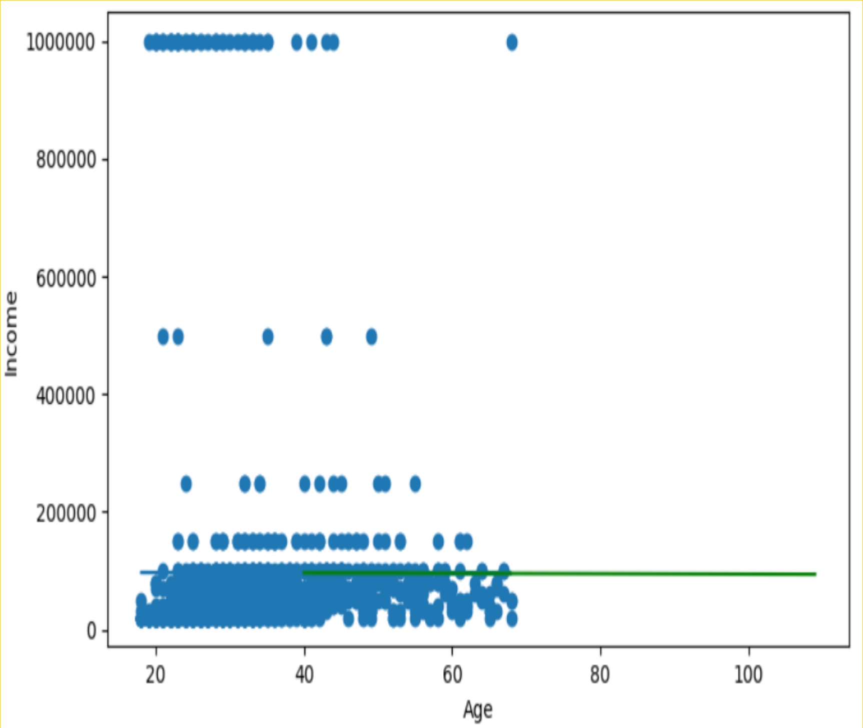
- Dataframe “attributes” was created with releant columns.
- A dictionary with keys being the preexisting column cell data and the value being the numerical representation of the categorical data.
- Pointing to the dataframe.value.map(dictionary), the replacement with the desired column values was able to happen for diet, drinks, smokes, drugs, and education.

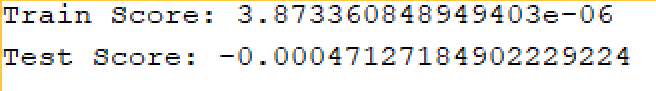


```
attributes = df[["diet", "drinks", "drugs", "body_type", "smokes",
"income", "age","education"]]
attributes["diet_code"] = attributes.diet.map(diet_type_map)
attributes["drinks_code"] = attributes.drinks.map(drink_type_map)
attributes["smokes_code"] = attributes.smokes.map(smoke_type_map)
attributes["drugs_code"] = attributes.drugs.map(drug_type_map)
attributes["body_code"] = attributes.body_type.map(body_type_map)
attributes["education_code"] =
attributes.education.map(education_type_map)
attributes.dropna(inplace=True)
attributes = attributes[attributes.income != -1]
attributes.drop(labels=["diet", "drinks", "drugs", "body_type",
"body_code", "smokes", "education"], axis=1, inplace=True)
feature_values = attributes.values
```

What dependency does age have on income?

- The data is very concentrated in various areas. The data is “weighted” based on which slope has the most weight”.
- The low “score” rate has likely to due with limitations of .fit(). Additionally, the correlation between age and income could be non-existent.
- The score() value illustrates that there is minimal correlation between income and height. So predicting this with a regression model would be very difficult.



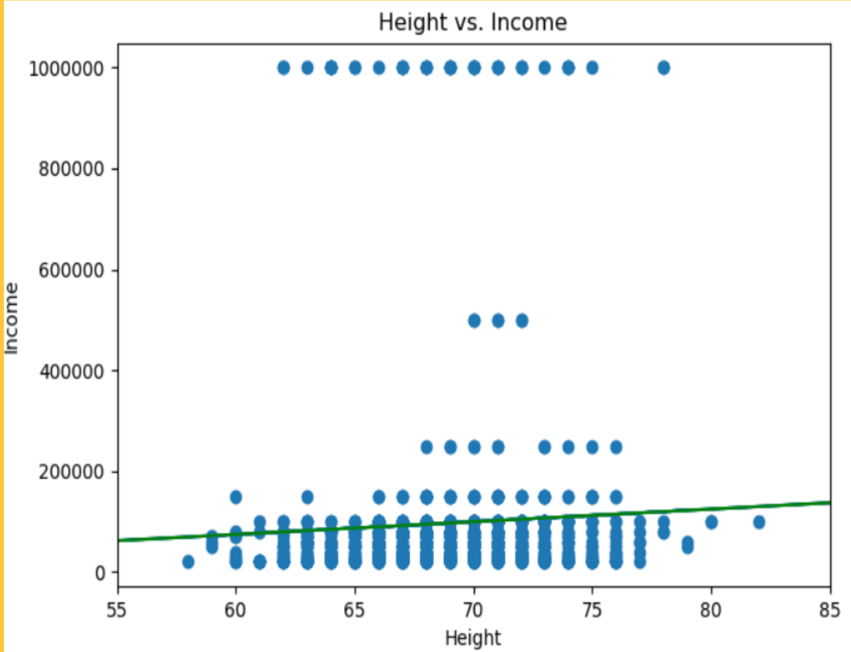


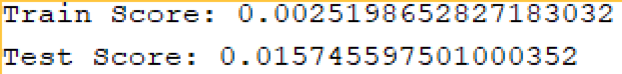
```
Train Score: 3.8733608489494O3e-06
Test Score: -0.00047127184902229224
```

codecademy

What dependency does height have on income?

- The data is very concentrated in various areas. The data is “weighted” based on which slope has the most weight”.
- The low “score” rate has likely to due with limitations of .fit(). Additionally, the correlation between height and income could be non-existent.
- The score() value illustrates that there is minimal correlation between income and height. So predicting this with a regression model would be very difficult.

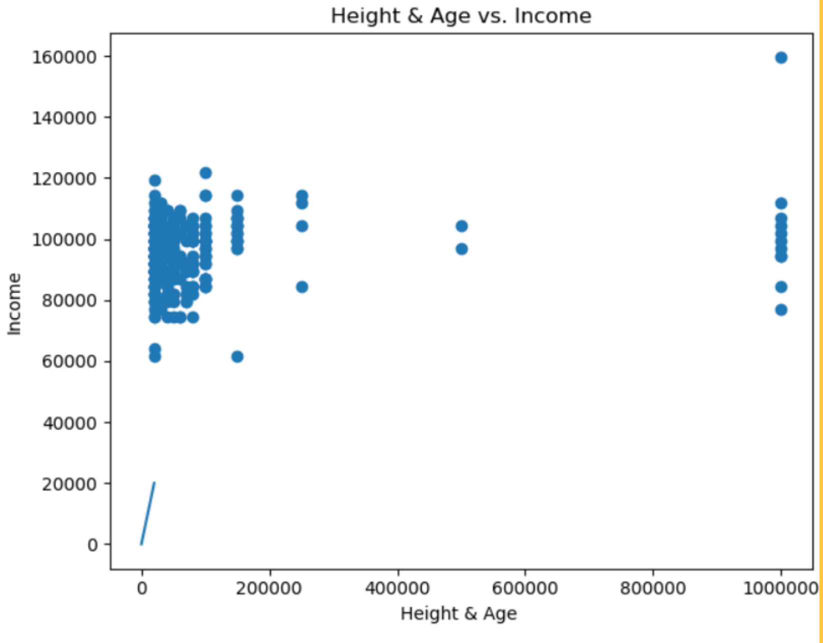


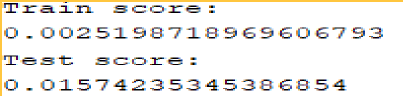


```
Train Score: 0.0025198652827183032
Test Score: 0.01574559750100352
```

What dependency does height & age have on income?

- The data is very concentrated in various areas. The data is "weighted" based on which slope has the most weight".
- The low "score" rate has likely to due with limitations of .fit(). Additionally, the correlation between height & age and income could be non-existent.
- The score() value illustrates that there is minimal correlation between income and height. So predicting this with a regression model would be very difficult.

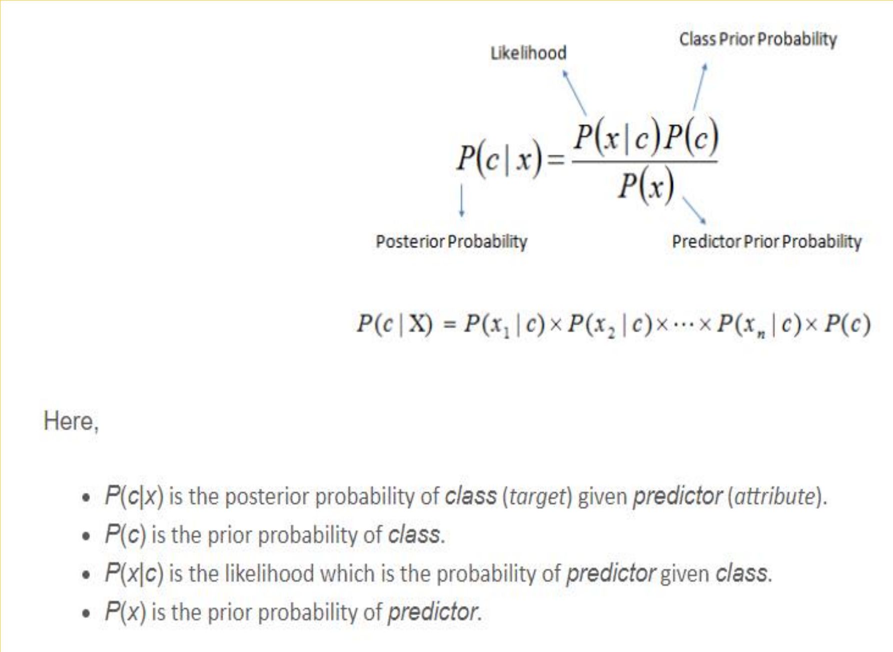




```
Train score:
0.0025198718969606793
Test score:
0.015742353453868854
```

What dependency does the frequency of smoking and drinking have on someone’s education level?

- A Naïve Bayes Classification is used here.
- This can be accomplished by doing the following:
    1. For each attribute (e.g. "graduated high school", "graduated college", "[smokes] rarely“, etc.) by using the using the .value_counts() method, count the amount of corresponding numerical values for each attribute in the dataframe.
    2. This relative frequency can be used as the “prior” for both the predictor and the class (see attached screenshot).
    3. For each user in test data, a “posterior probability” can be calculated to determine how to classify a particular user as having X education level GIVEN that they smoke AND drink.
- For this research question, it might be also helpful to use a regression model. Which will show the correlation between education and smoking and drinking.



$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|\mathrm{X}) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

Here,

- $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

Regression Approach

Here are some things I thought about:

Conclusion: The “.score()” values between the test data and predicted data are relatively small. This indicates a large lack of dependency between the income, age, and height.

Question: That question was whether it is possible to find the correlation between income, height, and age.

Next step: Select better independent variables (e.g. features) with better correlations with income than height and age. Additionally, trying other regression models with better correlations.

Classification Approach

Here are some things I thought about:

- **Naive Bayes is a good model to use for classification.**
- **However, a regression could also be used if what’s sought is a mere trend indicating dependency. If what’s sought it the ability to determine if someone smokes and drinks based on education, a classification (in this case Naïve Bayes) might be most appropriate.**

Conclusion and Next Steps

Here are some things I thought about:

Conclusion: The models are not very accurate or precise. It was also non-trivial mastering the syntax classes, methods, etc. for Pandas and numpy. If there is no correlation, there will likely be low or very erroneous values returned by .score() function. The preliminary answers to my asked questions are that there is no dependency.

Insight: Aberrant user data (e.g. astronomically high reported income levels, non-sensical replies, blanks, etc.) should be discarded in large datasets. Otherwise, aberrant data will skew the .score() returned values and reliability of the ML algorithm. Regressions usually work best for non-categorical data.

Next step: See what other ML algorithm (K- nearest neighbor, K-means Clustering, Naïve bayes Classification, other regressions) are suitable for answering correlations with income.

The End