Date-A-Scientist

By Ian De Bie

Table of Contents

Conclusion

- I ended up dropping NAN's from the dataset due to continually getting error when trying to fit the model.
- I don't believe the model was successful and would have been nice to not have so many NAN's
- This class was really complicated for me, but I am glad I took it in order to get a taste of what Machine Learning is all about

Exploring the Dataset

The first thing I did after creating Dataframe was to run head, describe, and value_count on religion. I noticed that the religion column wasn't grouped very well and if I wanted to use it then I would need to add a column revising it somewhat. Looking at the other columns gave me an idea what to start with.

	age	body_type	diet	...	smokes	speaks	status
0	22	a little extra	strictly anything	...	sometimes	english	single
1	35	average	mostly other	...	no	english (fluently), spanish (poorly), french (...	single
2	38	thin	anything	...	no	english, french, c++	available
3	23	thin	vegetarian	...	no	english, german (poorly)	single
4	29	athletic	NaN	...	no	english	single

	age	height	income
count	59946.000000	59943.000000	59946.000000
mean	32.340290	68.295281	20033.222534
std	9.452779	3.994803	97346.192104
min	18.000000	1.000000	-1.000000
25%	26.000000	66.000000	-1.000000
50%	30.000000	68.000000	-1.000000
75%	37.000000	71.000000	-1.000000
max	110.000000	95.000000	1000000.000000

agnosticism	2724
other	2691
agnosticism but not too serious about it	2636
agnosticism and laughing about it	2496
catholicism but not too serious about it	2318
atheism	2175
other and laughing about it	2119
atheism and laughing about it	2074
christianity	1957
christianity but not too serious about it	1952
other but not too serious about it	1554
judaism but not too serious about it	1517
atheism but not too serious about it	1318
catholicism	1064

Predict Religion?

- Can religion be predicted based off ethnicity, or off a combination of other columns?
- Try first just based off ethnicity, and then add diet, drugs, drinking, and smoking
- Need to modify religion, diet, and ethnicity to group them into simple categories
- Create numerical data from each of the columns to use as features

Column Creation

Religion, Diet, and Ethnicity columns contained many records with NAN. So I first wanted to change all of those to 'Other', with the reason being they didn't fill it out so it basically is other. The next challenge was to take all the different phrases and group it by the main category. For that part, I created a split function to take either the first or last word of each record depending on what I wanted to group by. After that, I took those 3 columns, plus drinks, drugs, and the smokes columns and converted those to a numeric code similar to the example in the Capstone.

```
df.religion = df.religion.fillna('other')
df.diet = df.diet.fillna('other')
df.ethnicity = df.ethnicity.fillna('other')

def split_col(data, index=0):
  output = str(data).split()
  return str(output[index]).strip(',')

df['diet_cat'] = df['diet'].apply(lambda x: split_col(x, -1))
df['religion_cat'] = df['religion'].apply(lambda x: split_col(x, 0))
df['eth_cat'] = df['ethnicity'].apply(lambda x: split_col(x, 0))
```

```
religion_mapping = {'other': 0, 'atheism': 1, 'agnosticism': 2,  'islam': 3, 'hinduism': 4, \
  'buddhism': 5, 'judaism': 6, 'catholicism': 7, 'christianity': 8}
diet_mapping = {'other': 0, 'halal': 1, 'kosher': 2, 'vegan': 3, 'vegetarian': 4, 'anything': 5}
eth_mapping = {'other': 0, 'native': 1, 'pacific': 2, 'middle': 3, 'indian': 4, 'black': 5, \
  'hispanic': 6, 'asian': 7, 'white': 8}
drinks_mapping = {'not at all': 0, 'rarely': 1, 'socially': 2, 'often': 3, 'very often': 4, \
  'desperately': 5}
drugs_mapping = {'never': 0, 'sometimes': 1, 'often': 2}
smokes_mapping = {'no': 0, 'sometimes': 1, 'when drinking': 2, 'trying to quit': 3, 'yes': 4}

df['religion_num'] = df.religion_cat.map(religion_mapping)
df['diet_num'] = df.diet_cat.map(diet_mapping)
df['eth_num'] = df.eth_cat.map(eth_mapping)
df['drinks_num'] = df.drinks.map(drinks_mapping)
df['drugs_num'] = df.drugs.map(drugs_mapping)
df['smokes_num'] = df.smokes.map(smokes_mapping)
```

Column Creation Continued...

Multiple Linear Regression