DataExplore

November 4, 2018

0.1 Project Problem and Hypothesis

- This project is to verify Angela Duckworth's study on "Grit" and attempt to expand on it
 with other features present in their dataset (http://angeladuckworth.com/). I will also set
 out to verify current scientific assumptions of the differences between genders and big five
 personality traits.
 - On average, individuals who are gritty are more self-controlled, but the correlation between these two traits is not perfect: Some individuals are paragons of grit but not self-control, and some exceptionally well-regulated individuals are not especially gritty (Duckworth & Gross, 2014)
 - Feel free to take the test! (http://angeladuckworth.com/grit-scale/) My score was 3.3 (average).
- This dataset has many features and I have the opportunity to try different hypothesis. My main focus will be to predict a "grit" score from personality traits and demographic data.
 - Specifically, Duckworth mentioned on the Freakonomics Podcast, that the big five trait
 conscientiousness as an existing success factor, and I'd like to see how similar grit and
 this trait are.
 - We can test some widely accepted personality traits by gender as well.

0.2 Caveats and Wishes

- This data is reflects people that took this online personality test survey. It may be biased toward people who are interested in how gritty they are (potentially people who are already gritty).
- Only users who were willing to share their data for research purposes are present in the data.
- Some things that would make this dataset even better would be the 10 aspects of the big 5, IQ, some sort of success criteria, and political affiliation. Additionally, if all survey takers took the same test a year later, we could find more results.

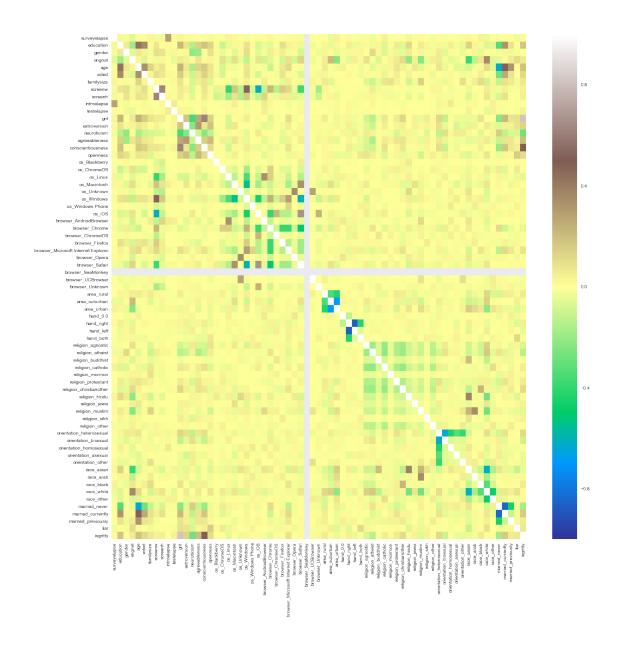
Neuroticism	Agreeableness	Conscientiousness	Extroversion	Openness/Intellect
Volatility	Compassion	Industriousness	Enthusiasm	1
Withdrawal	Politeness	Orderliness	Assertiveness	

```
In [281]: import pandas as pd
    import seaborn as sns
    import numpy as np
    from matplotlib import pyplot as plt
    %matplotlib inline
```

0.3 Load the data and remove a test set

0.4 Generate a heatmap of all of the features in the dataset

Out[292]: <matplotlib.axes._subplots.AxesSubplot at 0x12b37b5d0>



0.4.1 There's so much to see here!

Who are gritty people?

- Married, educated, older voters who are conscientious and agreeable without being prone to emotional anxiety.
- Internet explorer users tend to be grittier.

Big 5 and gender

- The biggest differences in the big 5 personality traits and gender are opennes, neuroticism, and agreeableness.
- We'll take a closer look at agreeableness.

Validating the data

- Older people tend to be married, educated, voter who have developed "positive" big 5 traits
- iOS users have smaller screens and use the Safari browser. Windows users have larger screens and often use the Chrome browser.
- Hindu and Buddhist survey takers tended to be Asian
- Muslim survey takers tended to be Arabian.

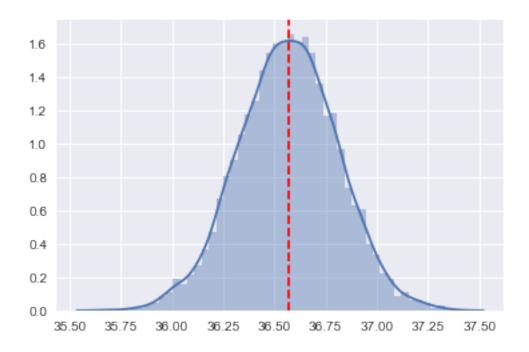
The "liar" (confidence/invalid) column

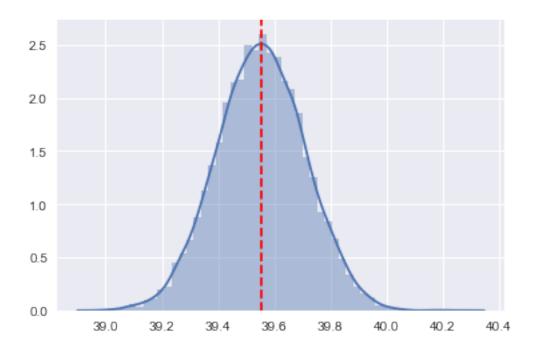
- Whites lied more often, Asians were more honest
- If you say you write with both hands, you're a liar
- · Liars answer more positively about themselves

```
In [558]: df['grit'].describe()
Out[558]: count
                    2211.000000
          mean
                       3.210274
          std
                       0.693705
          min
                       1.166667
          25%
                       2.666667
          50%
                       3.250000
          75%
                       3.750000
                       5.000000
          Name: grit, dtype: float64
```

1 Differences in agreeableness by gender

Let's bootstrap the mean from both groups and compare them to each other and the data in general.



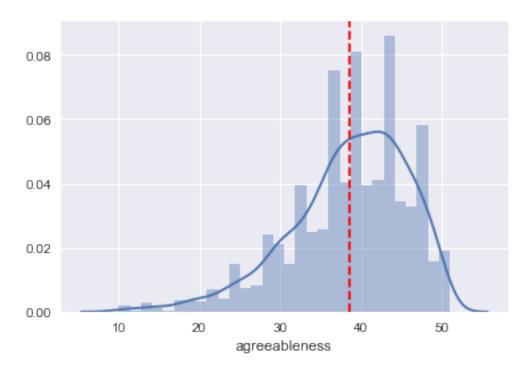


1.0.1 How confident are we that the means differ?

100% confident. There is no overlap in our male bootsrap samples and female bootstrap samples. We can keep this confidence up until about a difference of \sim 1.5

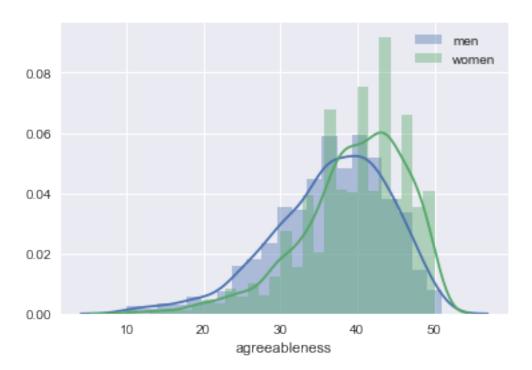
```
In [295]: count = 0
          men_max = np.max(men_agree_means)
          women_min = np.min(women_agree_means)
          print men_max
          print women_min
          print women_min - men_max
          for mean in women_agree_means:
              if mean < men_max:</pre>
                  count += 1
          print count
37.3986152324
38.978381096
1.57976586359
0
In [296]: print np.mean(df['agreeableness'])
          fig, ax = plt.subplots()
          sns.distplot(df['agreeableness'])
          ax.axvline(np.mean(df['agreeableness']),linestyle='--',color='red')
```

Out[296]: <matplotlib.lines.Line2D at 0x12a8e5850>



The mean of the bootsrap samples difference is just under a half of a standard deviation.

Out[297]: <matplotlib.legend.Legend at 0x12b079910>



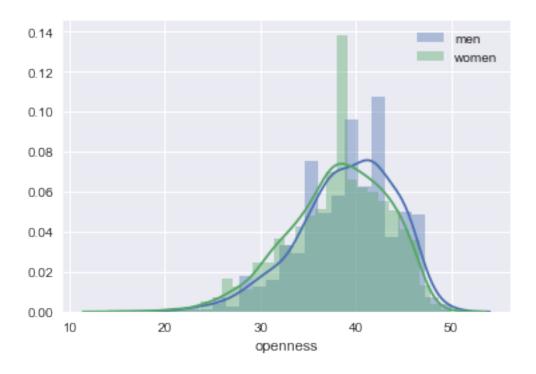
In [298]: df['agreeableness'].describe()

Out [298]: count 3000.000000 38.547667 mean 7.307452 std 10.000000 min 34.000000 25% 50% 40.000000 75% 44.000000 51.000000 max

Name: agreeableness, dtype: float64

1.1 Differences in openness?

• Present, but not as drastic as agreeableness



1.1.1 As suspected, grit is most highly correlated with conscientiousness:

```
In [300]: df['grit'].corr(df['conscientiousness'])
Out[300]: 0.61463512284440447
```

1.2 Train linear regression and decision tree models to describe the data

Linear regression with statsmodels (normalize to show feature importance):

```
In [546]: import statsmodels.api as sm
```

model = sm.OLS(dfY,dfX)
results = model.fit()
print results.summary()

OLS Regression Results

=======================================			=======================================
Dep. Variable:	isgritty	R-squared:	0.709
Model:	OLS	Adj. R-squared:	0.707
Method:	Least Squares	F-statistic:	446.4
Date:	Tue, 26 Sep 2017	Prob (F-statistic):	0.00
Time:	11:12:27	Log-Likelihood:	-1185.7
No. Observations:	2211	AIC:	2395.
Df Residuals:	2199	BIC:	2464.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
education	0.0286	0.011	2.556	0.011	0.007	0.051
gender	-0.0347	0.020	-1.770	0.077	-0.073	0.004
age	4.396e-05	0.001	0.051	0.959	-0.002	0.002
voted	-0.0039	0.021	-0.188	0.851	-0.045	0.037
familysize	0.0044	0.005	0.840	0.401	-0.006	0.015
extroversion	0.0016	0.001	1.518	0.129	-0.000	0.004
neuroticism	-0.0105	0.001	-10.992	0.000	-0.012	-0.009
agreeableness	0.0047	0.001	3.543	0.000	0.002	0.007
conscientiousness	0.0264	0.001	22.003	0.000	0.024	0.029
openness	-0.0038	0.002	-2.437	0.015	-0.007	-0.001
married_never	-0.1381	0.039	-3.578	0.000	-0.214	-0.062
married_currently	-0.0551	0.043	-1.293	0.196	-0.139	0.029

Kurtosis:	1.946	Cond. No.	499.
Skew:	-0.161	Prob(JB):	5.06e-25
Prob(Omnibus):	0.000	Jarque-Bera (JB):	111.886
Omnibus:	669.048	Durbin-Watson:	2.055

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

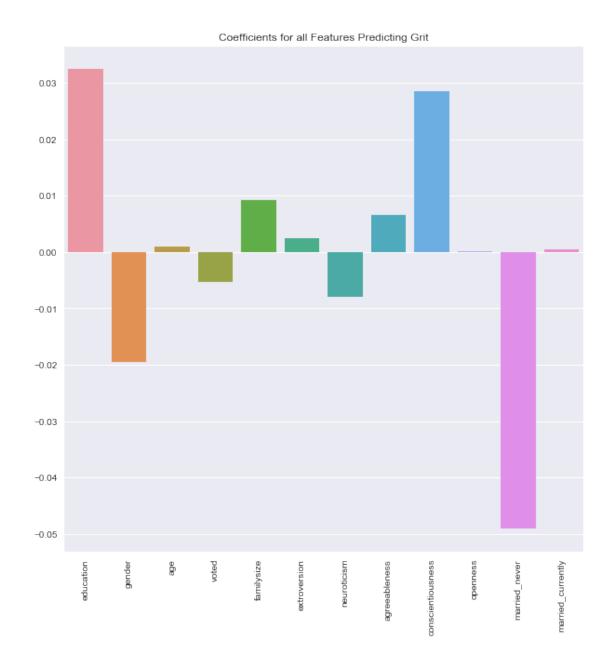
Linear regression with sklearn:

1.3 With the LR model trained, let's look at the coefficients and visualize the results

Keep in mind, education is ordinal

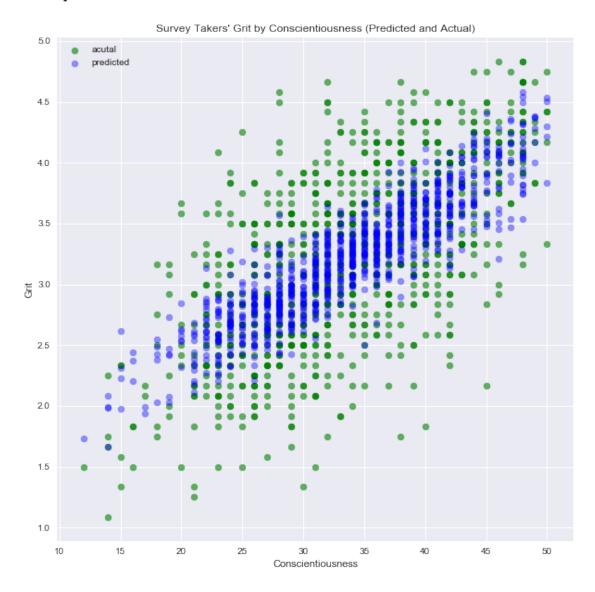
- Less than high school
- High school
- University degree
- Graduate degree

Gender and "liar" appear to have a big affect, but keep in mind the range of conscientiousness is 10-50

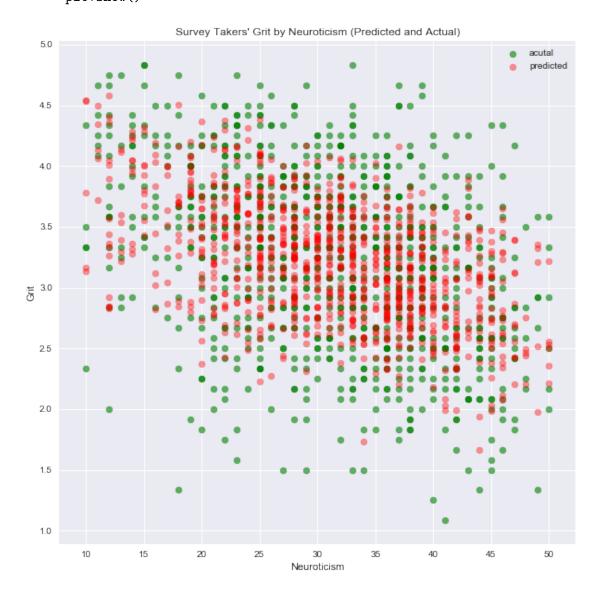


1.4 Let's visualize predictions in terms of conscientiousness and another feature (neuroticism)

```
plt.legend(["acutal","predicted"])
ax.set_xlabel('Conscientiousness')
ax.set_ylabel('Grit')
plt.show()
```



```
plt.title("Survey Takers' Grit by Neuroticism (Predicted and Actual)")
plt.legend(["acutal", "predicted"])
ax.set_xlabel('Neuroticism')
ax.set_ylabel('Grit')
plt.show()
```



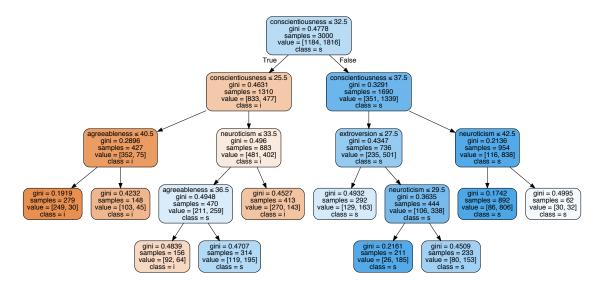
We can see that grit is highly correlated with conscientiousness. Our predictions vary only a little from the mean at each level of conscientiousness, while our predictions according to neuroticism have a much wider range.

1.5 Time to train and visualize a decision tree classifier

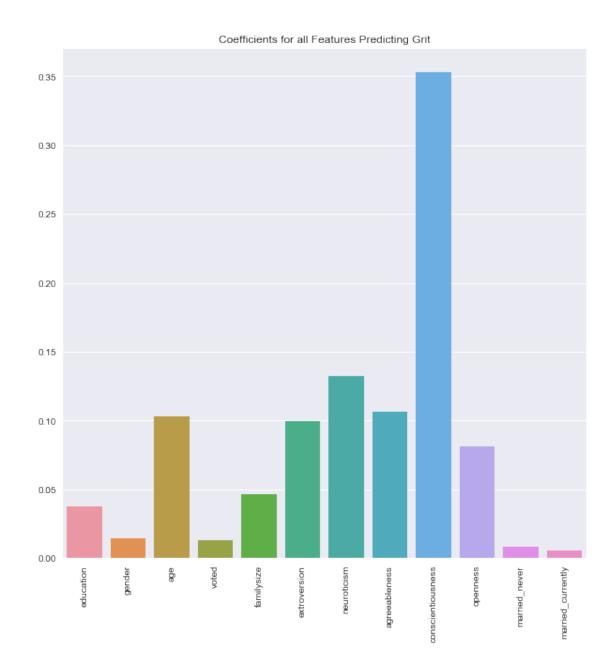
Something cool to keep in mind here is that the most important factors are personality traits. We could consider these as being an intrinsic cause for the other features as well

```
In [164]: from sklearn.tree import DecisionTreeClassifier
          from sklearn.model_selection import cross_val_score
          dfY = df['isgritty']
          dt = DecisionTreeClassifier(max_depth=5,max_leaf_nodes=10)
          print cross_val_score(dt, dfX, dfY, cv=10)
          mod = dt.fit(dfX,dfY,)
[ 0.70431894  0.77408638  0.71428571  0.66777409  0.73666667
                                                              0.76
  0.77257525 0.75585284 0.73913043 0.73578595]
In [165]: from sklearn import tree
          import graphviz
          dot_data = tree.export_graphviz(mod, out_file=None,
                                   feature names=indFeats,
                                   class_names='isgritty',
                                   filled=True, rounded=True,
                                   special_characters=True)
          graph = graphviz.Source(dot_data)
          graph
```

Out [165]:



1.6 If we train a random forest classifier, we see similar results in terms of which features are most important



1.7 We have a good sense of our data and the factors that relate to grit now, so let's train a nueron and a nueral network to predict grit

```
In [564]: from numpy import exp, array, random, dot

class Neuron():
    def __init__(self):
        #random.seed(1988)
        #start with random weights between -1 and 1
        self.weights = 2.0 * random.random((12,1)) - 1.0
```

```
#the sigmoid function is like the logistic function we used in logistic regressi
    #similarly, this allows for an output between 0 and 1
    #It looks like this is rarely used anymore in favor of the "RELU" function
    def __sigmoid(self, x):
        return 1 / (1 + np.exp(-x))
    #gradient of the sigmoid curve, this adjusts the severity of the error according
   def __sigmoid_derivative(self, x):
       return x * (1-x)
   def train(self, training_set_inputs, training_set_outputs, number_of_training_ite
        for iteration in xrange(number_of_training_iterations):
            #pass the training set through our neuron
            output = self.predict(training_set_inputs)
            #calculate the error
            error = training_set_outputs - output
            #multiply the error by the gradient of the sigmoid curve, then by the in
            adjustment = dot(training_set_inputs.T, error * self.__sigmoid_derivative
            #adjust the weights
            self.weights += adjustment
            #print
            #print training_set_inputs.T
    def predict(self, inputs):
        #pass inputs through our neuron
        return self.__sigmoid(np.array(np.dot(inputs, self.weights),dtype=np.float32
#initialise a single neuron neural network
neuron = Neuron()
print 'random starting weights:'
print neuron.weights
print
#train the neural network using a training set. do it 10,000 (or 600) times and mak
from sklearn.preprocessing import normalize
X = np.array(dfX.as_matrix(), dtype=float)
X = normalize(X, axis=0)#, norm='max')
dfY = df['isgritty']
Y = dfY.as_matrix().reshape(dfY.as_matrix().shape[0],-1)
neuron.train(X,Y,2000)
```

```
print 'new weights after training: '
          print neuron.weights
random starting weights:
[[ 0.54230874]
[ 0.15843374]
[ 0.81331905]
 [ 0.51774819]
 [ 0.38414109]
 [ 0.97296606]
 [ 0.93687881]
 [-0.2758979]
 [-0.95800895]
 [-0.27431323]
 [-0.70229964]
 [ 0.34965651]]
new weights after training:
[[ 13.58220469]
[-6.54035531]
 [-2.76242087]
 [ 0.45478189]
 [-2.74036934]
 [ 7.88566158]
 [-105.28036888]
 [ 25.67214587]
 [ 193.3658723 ]
[ -44.62970236]
 [ -58.51468447]
 [ -14.49317004]]
In [565]: print indFeats
          print('Coefficients: \n', lm.coef_)
['education', 'gender', 'age', 'voted', 'familysize', 'extroversion', 'neuroticism', 'agreeable
('Coefficients: \n', array([ 0.03245315, -0.01947561, 0.00098068, -0.00539197, 0.00915535,
        0.00239531, -0.00787071, 0.00653773, 0.02850376, 0.00017041,
       -0.04914387, 0.00048606]))
In [566]: from sklearn.metrics import accuracy_score
          #np.array(df_test_X.as_matrix(), dtype=float)
          X_test = normalize(np.array(df_test_X.as_matrix(), dtype=float), axis=0)#, norm='max
          Y_test = dfTest['isgritty']
          Y_test = Y_test.as_matrix().reshape(Y_test.as_matrix().shape[0],-1)
          print("Accuracy: %.2f"
                % accuracy_score(Y_test, np.round(neuron.predict(X_test))))
```

This single neuron behaves similarly to logistic regression We can train a neural net to take into account feature interaction (This is a bit of overkill)

```
In [567]: from keras.models import Sequential
     from keras.layers import Dense
     from keras.layers import Dropout
In [571]: # create model
     model = Sequential()
     model.add(Dense(32,activation='relu',input_shape=(12,)))
     model.add(Dropout(0.3))
     model.add(Dense(16,activation='relu'))
     model.add(Dense(8,activation='relu'))
     model.add(Dense(4,activation='relu'))
     model.add(Dense(1,activation='sigmoid'))
      # Compile model
     model.compile(loss='binary_crossentropy', optimizer='rmsprop', metrics=['accuracy'])
In [578]: import datetime
     print datetime.datetime.now()
      # Fit the model
     model.fit(dfX.as_matrix(),dfY.as_matrix(), epochs=300, batch_size=200)
     print datetime.datetime.now()
2017-09-26 12:29:44.838318
Epoch 1/300
Epoch 2/300
Epoch 3/300
Epoch 4/300
Epoch 5/300
Epoch 6/300
Epoch 7/300
Epoch 8/300
```

```
Epoch 9/300
Epoch 10/300
Epoch 11/300
Epoch 12/300
Epoch 13/300
Epoch 14/300
Epoch 15/300
Epoch 16/300
Epoch 17/300
Epoch 18/300
Epoch 19/300
Epoch 20/300
Epoch 21/300
Epoch 22/300
Epoch 23/300
Epoch 24/300
Epoch 25/300
Epoch 26/300
Epoch 27/300
Epoch 28/300
Epoch 29/300
Epoch 30/300
Epoch 31/300
Epoch 32/300
```

```
Epoch 33/300
Epoch 34/300
Epoch 35/300
Epoch 36/300
Epoch 37/300
Epoch 38/300
Epoch 39/300
Epoch 40/300
Epoch 41/300
Epoch 42/300
Epoch 43/300
Epoch 44/300
Epoch 45/300
Epoch 46/300
Epoch 47/300
Epoch 48/300
Epoch 49/300
Epoch 50/300
Epoch 51/300
Epoch 52/300
Epoch 53/300
Epoch 54/300
Epoch 55/300
Epoch 56/300
```

```
Epoch 57/300
Epoch 58/300
Epoch 59/300
Epoch 60/300
Epoch 61/300
Epoch 62/300
Epoch 63/300
Epoch 64/300
Epoch 65/300
Epoch 66/300
Epoch 67/300
Epoch 68/300
Epoch 69/300
Epoch 70/300
Epoch 71/300
Epoch 72/300
Epoch 73/300
Epoch 74/300
Epoch 75/300
Epoch 76/300
Epoch 77/300
Epoch 78/300
Epoch 79/300
Epoch 80/300
```

```
Epoch 81/300
Epoch 82/300
Epoch 83/300
Epoch 84/300
Epoch 85/300
Epoch 86/300
Epoch 87/300
Epoch 88/300
Epoch 89/300
Epoch 90/300
Epoch 91/300
Epoch 92/300
Epoch 93/300
Epoch 94/300
Epoch 95/300
Epoch 96/300
Epoch 97/300
Epoch 98/300
Epoch 99/300
Epoch 100/300
Epoch 101/300
Epoch 102/300
Epoch 103/300
Epoch 104/300
```

```
Epoch 105/300
Epoch 106/300
Epoch 107/300
Epoch 108/300
Epoch 109/300
Epoch 110/300
Epoch 111/300
Epoch 112/300
Epoch 113/300
Epoch 114/300
Epoch 115/300
Epoch 116/300
Epoch 117/300
Epoch 118/300
Epoch 119/300
Epoch 120/300
Epoch 121/300
Epoch 122/300
Epoch 123/300
Epoch 124/300
Epoch 125/300
Epoch 126/300
Epoch 127/300
Epoch 128/300
```

```
Epoch 129/300
2211/2211 [============== ] - 0s - loss: 0.5313 - acc: 0.7404
Epoch 130/300
Epoch 131/300
Epoch 132/300
Epoch 133/300
Epoch 134/300
Epoch 135/300
Epoch 136/300
Epoch 137/300
Epoch 138/300
Epoch 139/300
Epoch 140/300
Epoch 141/300
Epoch 142/300
Epoch 143/300
Epoch 144/300
Epoch 145/300
Epoch 146/300
Epoch 147/300
Epoch 148/300
Epoch 149/300
Epoch 150/300
Epoch 151/300
Epoch 152/300
```

```
Epoch 153/300
Epoch 154/300
Epoch 155/300
Epoch 156/300
Epoch 157/300
Epoch 158/300
Epoch 159/300
Epoch 160/300
Epoch 161/300
Epoch 162/300
Epoch 163/300
Epoch 164/300
Epoch 165/300
Epoch 166/300
Epoch 167/300
Epoch 168/300
Epoch 169/300
Epoch 170/300
Epoch 171/300
Epoch 172/300
Epoch 173/300
Epoch 174/300
Epoch 175/300
Epoch 176/300
```

```
Epoch 177/300
Epoch 178/300
Epoch 179/300
Epoch 180/300
Epoch 181/300
Epoch 182/300
Epoch 183/300
Epoch 184/300
Epoch 185/300
Epoch 186/300
Epoch 187/300
Epoch 188/300
Epoch 189/300
Epoch 190/300
Epoch 191/300
Epoch 192/300
Epoch 193/300
Epoch 194/300
Epoch 195/300
Epoch 196/300
Epoch 197/300
Epoch 198/300
Epoch 199/300
Epoch 200/300
```

```
Epoch 201/300
2211/2211 [============== ] - 0s - loss: 0.5376 - acc: 0.7395
Epoch 202/300
Epoch 203/300
Epoch 204/300
Epoch 205/300
Epoch 206/300
Epoch 207/300
Epoch 208/300
Epoch 209/300
Epoch 210/300
Epoch 211/300
Epoch 212/300
Epoch 213/300
Epoch 214/300
Epoch 215/300
Epoch 216/300
Epoch 217/300
Epoch 218/300
Epoch 219/300
Epoch 220/300
Epoch 221/300
Epoch 222/300
Epoch 223/300
Epoch 224/300
```

```
Epoch 225/300
Epoch 226/300
Epoch 227/300
Epoch 228/300
Epoch 229/300
Epoch 230/300
Epoch 231/300
Epoch 232/300
Epoch 233/300
Epoch 234/300
Epoch 235/300
Epoch 236/300
Epoch 237/300
Epoch 238/300
Epoch 239/300
Epoch 240/300
Epoch 241/300
Epoch 242/300
Epoch 243/300
Epoch 244/300
Epoch 245/300
Epoch 246/300
Epoch 247/300
Epoch 248/300
```

```
Epoch 249/300
Epoch 250/300
Epoch 251/300
Epoch 252/300
Epoch 253/300
Epoch 254/300
Epoch 255/300
Epoch 256/300
Epoch 257/300
Epoch 258/300
Epoch 259/300
Epoch 260/300
Epoch 261/300
Epoch 262/300
Epoch 263/300
Epoch 264/300
Epoch 265/300
Epoch 266/300
Epoch 267/300
Epoch 268/300
Epoch 269/300
Epoch 270/300
Epoch 271/300
Epoch 272/300
```

```
Epoch 273/300
Epoch 274/300
Epoch 275/300
Epoch 276/300
Epoch 277/300
Epoch 278/300
Epoch 279/300
Epoch 280/300
Epoch 281/300
Epoch 282/300
Epoch 283/300
Epoch 284/300
Epoch 285/300
Epoch 286/300
Epoch 287/300
Epoch 288/300
Epoch 289/300
Epoch 290/300
Epoch 291/300
Epoch 292/300
Epoch 293/300
Epoch 294/300
Epoch 295/300
Epoch 296/300
```

2 Conclusions

- Grit is highly correlated with conscientiousness (or conscientiousness is a great predictor of grit). It may be a redefinition of industriousness.
- A person's natural self (big 5 personality traits) is a better predictor of grit than other life outcomes.
- There are differences between some big 5 traits across demographics.
- For simple data sets, a single neuron can predict outcomes just as well as a neural network.

In []: