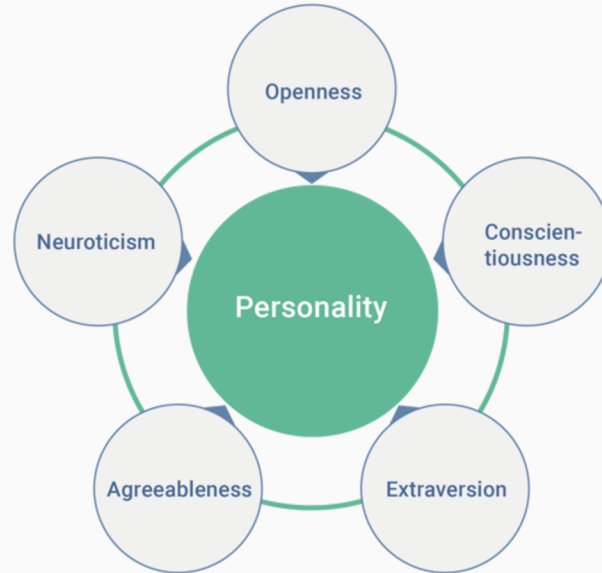


Machine Learning in Practice:
The Big 5 Personality Test

By: Elliot Allard & Brandyn Amado

For the class: Practical Data Science, S20
Raja Sooriamurthi

In psychology, the five factor model is a well-accepted taxonomy for grouping personality traits.



In psychology, the five factor model is a well-accepted taxonomy for grouping personality traits.

- **Openness**
cautious *Inventive vs.*
- **Extroversion**
solitary *Outgoing vs.*
- **Conscientiousness** *Efficient vs. easy-going*
- **Agreeableness**
challenging *Friendly vs.*

Our dataset is taken from the Open-Source Psychometrics Project, which hosts a free, online FFM test.

- Found on Kaggle: <https://www.kaggle.com/tunguz/big-five-personality-test>
- Over 1M rows, or responses to the test (great for ML)
- 110 Columns (mostly question responses)
- Manually calculated aggregate “scores” for each personality trait

Our Questions:

Can you predict someone's country of origin, given their totals in each factor?

- Based off of this [article](#), the FFM does not hold up (there was not a huge distinction between factors) when tested outside of urban, literate populations
- Attempted logistic regression with multiple input columns
- Found our accuracy to be around the same as randomly choosing a country

Can you predict someone's fifth personality score given the other four?

- Wanted to see if these personality areas were related in any way
- Same result for each score: around 50% accuracy of prediction
- Much better than guessing, but not too much influence over the score
- Tried a variety of regression models
- This speaks to the soundness of the FFM

Can you predict someone's answer to a question, given their previous answers in the same category?

- Wanted to see if the questions overlapped at all
- Same result for each question: around 50% accuracy of prediction
- Again, much better than guessing (which would give 20% accuracy)
- Tried a variety of regression models
- This speaks to the breadth of the questions, they don't seem to overlap

Our Conclusions:

Little redundancy was found within the model and the survey questions

- The five factor model of personality shows some relation between personality areas, though not drastically
- Gives credibility to the independence of factors in the model
- Question responses are somewhat correlated
- Speaks to the lack of redundancy in the survey

But, our dataset does have limitations.

- The large majority of respondents have IP addresses located in either North America or the EU
- IP addresses are also an unreliable way of locating country - there are many methods of spoofing them
- Looking at the barriers to taking this - having a computer, being interested in your FFM results

Thank you!



Machine Learning in Practice:
The Big 5 Personality Test

By: Elliot Allard & Brandyn Amado

For the class: Practical Data Science, S20
Raja Sooriamurthi