Predicting personality through the Big Five using data mining concepts

Chengpneg Wu Department of Computer ScienceGeorgia State UniversityAtlanta, GA 30303

*Abstract*—Scientists of humanity and sociology had already defined five characteristics of different individuals commonly known as big five characteristics namely, openness, neuroticism, conscientiousness, agreeableness and extraversion. By detecting 7 different factors of huge dataset from scientific survey. These 5 characteristics of different individuals are stored in a dataset and used for training. The goal of this project is using data mining concepts and python data science libraries to analyze and classify personalities of a given set of people or an individual.

Keywords—personality, prediction, linear model, random forest, data mining

# Introduction

For so many years study and different research by the humanity scientists. The Big Five personality was determined in nowadays. The 5 different personality traits are extraversion (also often spelled extroversion), agreeableness, openness, conscientiousness, and neuroticism. Each trait represents a continuum. Individuals can fall anywhere on the continuum for each trait [1]. These Big Five personality traits can be simply used to predict different human personality from different scoring such as serious, lively, responsible, and so on.

## Gender

Up on the passing centuries, women were more than often scored higher on the Five Factor Model (The Big Five) traits of Neuroticism and Agreeableness. [2] The first one represents painful proneness and high percentage of experiencing of a variety of negative effects. Meanwhile the second one reflects more positive parts such as kind, altruism, trust, tenderness, and compliance. Gender differences are taking huge effect on the personality forming: Costa’s report showed worldwide adult women scored .51 SD higher on Neuroticism and .59 SD higher on Agreeableness than man does. Which also revels the high-quality affections of gender differences to personality predictions.

## Age

Roberts BW had found that the average levels of Social Vitality tended to be commonly stable through the lifespan and it reached peak of the model in the early 20s age. Meanwhile this study also found that amount the ages between mid-20s till mid-50s, there was obviously but slightly decline of the Social Vitality [3]. Meanwhile age differences in the Big Five have also been corrected in cross-cultural research [4]. McCrae used the simple samples from Croatia, Portugal, Italy, Germany, and Korea and determine that Extraversion and Openness are much lower in old age participants than the youngers. Meanwhile the Conscientiousness and Agreeableness are showing the exact opposite pattern. [5]

## The Big Five

OPENNESS

The word "openness" should be used to explain something new or experiential, such as acceptance. In terms of openness, people with high scores are more curious and imaginative than those with low scores, which reflects their openness to new ideas. It shows how enlightened a person is. Highly open individuals enjoy thinking and learning, are sensitive to art and beauty, and generate original ideas, whereas close-minded individuals tend to have a narrow range of intellectual and creative interests [6].

NEUROTICISM

Neuroticism is a trait tendency to experience negative effects, including anger, anxiety, self-awareness, irritability, emotional instability, and depression. It captures differences in the frequency and intensity of negative emotions. Highly neurotic individuals are prone to experiencing anxiety, sadness, and mood swings, whereas emotionally stable individuals tend to remain calm and resilient, even in difficult circumstances [6].

CONSCIENTIOUSNESS

Conscientiousness is defined as "the tendency to react in a particular way in a particular situation," or more generally, the tendency to think, feel and act in a relatively persistent and consistent way over a period of time, while providing a trait. Conscientiousness represents differences in organization, productiveness, and responsibility. Highly conscientious individuals prefer order and structure, work persistently to pursue their goals, and are committed to fulfilling their duties and obligations, whereas unconscientious individuals are comfortable with disorder and less motivated to complete tasks [6].

AGREEABLENESS

Agreeableness is a common characteristic of amiability. Generally speaking, agreeableness refers to a person's ability to put other people's needs above his own. For example, those who are amiable are naturally compassionate and often enjoy serving and caring for others. Agreeableness captures differences in compassion, respectfulness, and acceptance of others. Agreeable individuals experience emotional concern for others’ well-being, treat others with regard for their personal rights and preferences, and hold generally positive beliefs about others; disagreeable individuals tend to have less regard for others, and for social norms of politeness [6].

EXTRAVERSION

Its characteristic is that a person's interest and energy tend to the external world of people and things, rather than the internal world of subjective experience. e, Extraversion represents individual differences in social engagement, assertiveness, and energy level. Highly extraverted individuals enjoy socializing with others, are comfortable expressing themselves in group situations, and frequently experience positive emotions such as enthusiasm and excitement; in contrast, introverted individuals tend to be socially and emotionally reserved [6].

# Model Approach

Based on the introduction, there would be total 7 factors that affecting the result of the personality prediction: Gender, Age, The big five personality traits. At this point, the goal of this this project is to determine the best prediction through the most accurate and fittest model.

## Choosing Linear Models

Since the number of factors are multiple, and the result that we trying to predict is simply just one. The linear model would be the best fit at this point. The following expression is a set of methods intended for regression in which the target value is expected to be a linear combination of the features. In mathematical notation if ȳ is the predicted value [8]. The mathematical equation is shown below as Fig1.0.

Chart, box and whisker chart

Description automatically generated

Fig1.0

Across the module, we designate the vector ω = (ω0 + ω1 + … + ωn) as coef\_ and ωn as intercept\_. However, in order to perform classification with generalized linear model, we will need to use Logistic regression.

## Logistic Regression

Speaking about logistic regression, even though its name is “regression”, however, it is a linear model as a matter of fact. Logistic regression is usually used for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier [9]. Basically, it models the probabilities that describe the possible outcomes of a single trial using a logistic function. Which is what we need in our project in order to predict the possible outcomes from the dataset.

## Logistic function

A logistic function or logistic curve is a common S-shaped curve (sigmoid curve) with equation below shown as Fig1.1:

Diagram

Description automatically generated with low confidence

Fig1.1

Where x0 represent the x value of the sigmoid's midpoint; L represent the curve's maximum value and k represent the logistic growth rate or steepness of the curve [10].

And one of the standard logistic sigmoid function where L = 1, k = 1, x0 = 0. It’s graph is shown below as Fig2.0.

Chart, histogram

Description automatically generated

Fig2.0

## Back to Logistic Regression

Logistic regression is implemented in LogisticRegression in “sklearn” python language environment. As its easy approaching, it can fit binary, One-vs-Rest, or multinomial logistic regression with optional l1, l2 or Elastic-Net regularization., or Elastic-Net regularization [8].

As an optimization, binary class l2 penalized logistic regression minimizes the cost function shown below as Fig3.1 [9]:

Text

Description automatically generated

Fig3.1

Similarly, l1 regularized logistic regression solves the following optimization pattern shown below as Fig3.2 [9]:

Text

Description automatically generated

Fig3.2

Elastic-Net regularization is a combination of l1 and l2 and minimizes the following cost as Fig3.3 [9]:

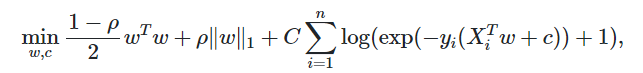


Fig3.3

Where p controls the strength of l1 regularization vs. l2 regularization (it corresponds to the l1\_ratio parameter).

The solver “liblinear” uses a coordinate descent (CD) algorithm, and relies on the excellent C++ LIBLINEAR library, which is shipped with scikit-learn [7]. However, the CD algorithm implemented in liblinear cannot learn a true multinomial (multiclass) model; instead, the optimization problem is decomposed in a “one-vs-rest” fashion so separate binary classifiers are trained for all classes [9].

# Prepare Experiment

The goal of this project is using data mining concepts and python data science libraries to analyze and classify personalities of a given set of people or an individual. This project will come across areas where it has access to large amounts of person behavioral data. This project will first train from the given datasets form a .csv file which contains 700 different data and their defined personality type as the project reference.

## Dataset Description

This project uses survey data from the website Kaggle for big five traits test : <https://www.kaggle.com/tunguz/big-five-personality-test>. The dataset includes several data for example: Age, Gender and so on, with level form 1 – 8 and. In order to make efficiency and more accuracy predictions, the range of tested age is from 17 – 28, which falls in the range of 1-10 along with other BIG FIVE traits test result too.

## Flow Chart

The process includes dataset creation, data pre-processing, data exploration, modeling, evaluation and interpretation as showed in Fig4.0.

Diagram

Description automatically generated

Fig4.0

## Data type

The collected data can be classified as table shown below as tab1.0:

|  |  |  |  |
| --- | --- | --- | --- |
| S.NO | ATTRIBUTE | TYPE | RANGE |
| 1 | Gender | nominal | Male/Female |
| 2 | Age | numeric | 17-28 |
| 3 | Openness | numeric | 1-8 |
| 4 | Neuroticism | numeric | 1-8 |
| 5 | Conscientiousness | numeric | 1-8 |
| 6 | Agreeableness | numeric | 1-8 |
| 7 | extraversion | numeric | 1-8 |

tab1.0

As it shown, the main goal of the preprocessing should be converting the nominal data type “Gender” to numeric data type. In this case, I decided to let “0” represent male and “1” represent female. And after this preprocessing, the data chart is shown as below as tab1.1. The data type would all be same as numeric, which is easy and highly efficient for generating the prediction.

|  |  |  |  |
| --- | --- | --- | --- |
| S.NO | ATTRIBUTE | TYPE | RANGE |
| 1 | Gender | nnumeric | 0-1 |

tab1.1

# Experimental Results

Given test dataset contains 8 attributes. Besides the 7 prediction references, the 8th one would be the identified personality for us to maintain the accuracy check.

The sample test table tab2.1 is given below from Kaggle website:

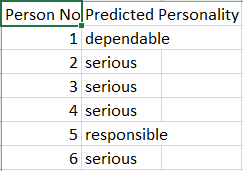
Scatter chart

Description automatically generated with medium confidence

tab2.1

While the whole prediction process will be using the only first 7 columns as references.

Linear model can provide the result with high accuracy of 0.85. The sample test result table tab2.2 is given after the project finished prediction:

The sample given results just doing the prediction for the first 6 person in the list. The whole list will be given in the attached zip file.

tab2.2

# Not used model - Random Forest Model

At the beginning of this project, I firstly choose the different model from the linear model, which is random forest model. However, this kind of model does not generate high accuracy as expected.

The result that I got from the random forest model is only have 0.3 accuracy. And the sample result is given as below as tab2.3 for the same test dataset:

Table

Description automatically generatedAs we can see clearly, the output(prediction) is quite different from the given identified personalities.

tab2.3

The reason for the inaccuracy can be found in the algorithm of the random forest model. Random forest, as the name suggests, is composed of a large number of individual decision trees, which run as a whole. Each tree in the random forest will throw out a class prediction, and the class with the most votes will become the prediction of our model.

The figure 4.0 clearly showed how the model is used to generalize part of the datasets. However, in our project random forest seems hard to generalize the given sets from the partial datasets. Which leading the prediction to the totally different way as the result.

Diagram

Description automatically generated

Fig4.0

##### Acknowledgment

I would like to thank .

##### References

1. Cherry, K. (n.d.). What are the big 5 personality traits? Retrieved April 15, 2021, from https://www.verywellmind.com/the-big-five-personality-dimensions 2795422#:~:text=The%20five%20broad%20personality%20traits,openness%2C%20conscientiousness%2C%20and%20neuroticism.
2. Costa PT, Terracciano A, McCrae RR. Gender differences in =pPersonality traits across cultures: Robust and surprising findings. Journal of Personality and Social Psychology. 2001;81:322–331. [PubMed] [Google Scholar]
3. Roberts BW, Walton KE, Viechtbauer W. Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. Psychological Bulletin. 2006;132:1–25. [PubMed] [Google Scholar]
4. McCrae RR, Costa PT, Pedroso de Lima, Simões A, Ostendorf F, Angleitner A, et al. Age differences in personality across the adult life span: Parallels in five cultures. *Developmental Psychology.*1999;35:466–477. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/10082017)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Developmental+Psychology&title=Age+differences+in+personality+across+the+adult+life+span:+Parallels+in+five+cultures.&author=RR+McCrae&author=PT+Costa&author=+Pedroso+de+Lima&author=A+Sim%C3%B5es&author=F+Ostendorf&volume=35&publication_year=1999&pages=466-477&pmid=10082017&)]
5. Helene H. Fung, PhD, Aging in Culture, The Gerontologist, Volume 53, Issue 3, June 2013, Pages 369–377, <https://doi.org/10.1093/geront/gnt024>
6. Soto, C. J. ―Big Five personalities traits.‖ ResearchGate, PP.240-241, 2018
7. “Plot Ridge Coefficients as a Function of the Regularization¶.” Scikit, scikit-learn.org/stable/auto\_examples/linear\_model/plot\_ridge\_path.html.
8. 1.1. linear Models¶. (n.d.). Retrieved April 17, 2021, from <https://scikit-learn.org/stable/modules/linear_model.html>
9. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
10. Logistic function. (2021, April 5). In Wikipedia. https://en.wikipedia.org/wiki/Logistic\_function