Gender Classification Based on a Five-Factor Model Personality Test

Abstract

In academic psychology, the Five-Factor Model is a group of five personality traits (Extraversion, Neuroticism, Agreeableness, Conscientiousness and Openness to Experience) that is thought to describe most of the differences in personality across individuals. This study will examine if the gender of an individual can be determined based on their answers to a 50-question Likert questionnaire that uses the Big-Five Factor Markers from the International Personality Item Pool (IPIP).

Through this, it is hoped that a tool can be devised that can help to create surveys that ask the respondent for no identifying information, with the goal of increasing response rates and honesty in responses, yet can still extract important demographic information about the individual.

Literature Review

In 1990, Goldberg examined the hypothesis that the personality traits among individuals that are the most important in human interaction will eventually come to be encoded in natural language. Using a factor analysis of English trait adjectives, it was determined that the five traits mentioned above capture the most important differences in personality and that other traits can be explained in terms of these five (Goldberg 1990).

Since the introduction of the Five-Factor Model, many studies have explored the relationship between personality and several demographic categories. For example, it has been seen that women tend to score higher in the areas of Neuroticism, Extraversion, Agreeableness and Conscientiousness than men, with Neuroticism tending to be the most marked difference. It has also been seen that gender differences are more prominent in countries with greater prosperity (Costa et al 2001).

On the subject of survey design, studies have shown that a survey that is anonymous can produce very different results than one that is simply confidential. One study found that on an anonymous survey, 74% of students that were known to have cheated admitted to doing so, while only 25% admitted to cheating through the confidential survey. Other questions on the survey also showed dramatic differences in responses (Ong et al 2006).

Other studies have shown that people have reported lower social anxiety and social desirability and higher self-esteem when they completed an anonymous survey compared to a non-anonymous version. (Joinson 1999).

Dataset

The dataset was collected (c. 2012) through an interactive online personality test on the website personality-testing.info and can be found at http://personality-testing.info/_rawdata/BIG5.zip. Participants were informed that their responses would be recorded and used for research at the beginning of the test and asked to confirm their consent at the end of the test.

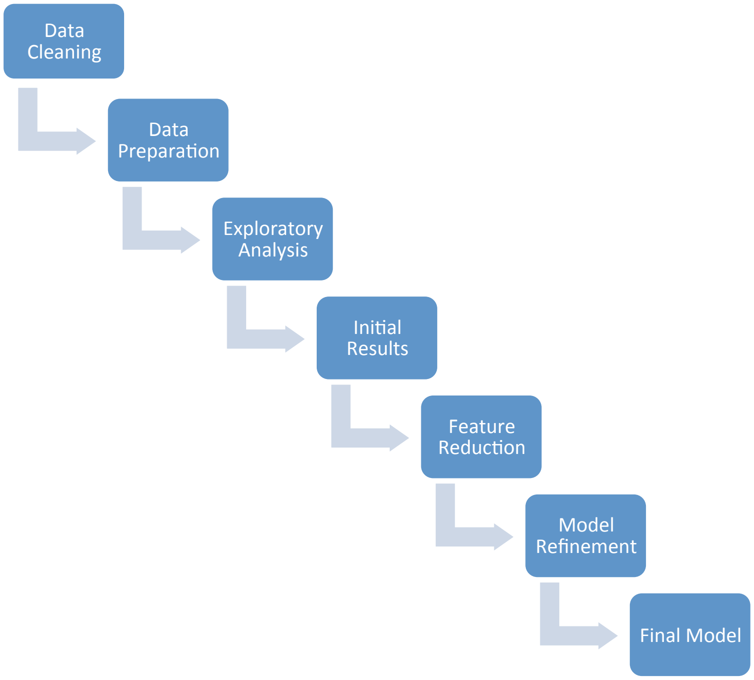
The dataset consists of 19719 observations and 57 variables. The variables are:

- race: Chosen from a drop down menu with 13 options.
- age: Entered as text.
- engnat: Yes/No response to "Is English your native language?"
- gender: Chosen from a drop down menu with 3 options (Male/Female/Other).
- hand: Right/Left/Both response to "What hand do you use to write with?"
- source: How the participant came to the test based on HTTP referer. (6 options)
- country: The ISO country code for the participant's technical location.
- 50 responses rated on a 5-point scale (1=Disagree, 3=Neutral, 5=Agree). The variable names are coded using a letter representing the personality trait tested (E, N, A, C, O) and a number between 1 and 10.

All missing values are coded as 0.

The variables that will be used in the study are all 50 questionnaire responses and gender.

Approach



Step 1: Data Cleaning

R Script: https://github.com/davidlichacz/Capstone_Project/blob/master/dataclean.R

The data was cleaned with the following steps:

- The columns race, age, engnat, hand, source and country were removed as they will not be used in the analysis.
- The only missing values in the questionnaire data occur in row 19605 in which all 50 values are missing. As a result, this row was discarded.
- There were 24 missing values in the gender variable. These were replaced using Hot Deck Imputation, a popular method for missing survey data.
- There were 102 values in the gender variable labeled as Other. Since the analysis will be concerned with classifying as Male or Female, these were also discarded.

The final data frame that will be used for analysis consists of 19606 rows and 59 columns.

Step 2: Data Preparation

R Script: https://github.com/davidlichacz/Capstone_Project/blob/master/datapreparation.R

The data was prepared for exploratory analysis with the following steps:

- Male and Female factors were created for the gender column.
- A new data frame was created that stores the total score of each individual for each personality trait. Scores were calculated based on the formulas used by personality-testing.info.
- Since the questions for each personality trait are a mixture of positively and negatively phrased questions, the negative questions were transformed to be positive so that graphical comparison is possible. This was accomplished with a simple linear transformation resulting in no loss of information.
- A new data frame was created that stores three new variables (Extreme, Moderate, Neutral) that contain counts of the number of extreme Likert values (1 or 5) chosen by an individual, the number of moderate values (2 or 4) and the number of neutral values (3).

Step 3: Exploratory Analysis

R Scripts: https://github.com/davidlichacz/Capstone_Project/blob/master/likertgraphs.R

https://github.com/davidlichacz/Capstone_Project/blob/master/exploratory.R

Six Likert graphs were produced, one that compares all 50 questions in the survey and one for each of the five personality traits. The graphs show similarities in the distribution for some pairs of questions suggesting that some feature reduction will be beneficial and that a shorter and more efficient questionnaire could be produced with little to no loss of information.

Plots:

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/likert_allquestions.jpeg

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/likert_agreeableness.jpeg

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/likert_conscientiousness.jpeg

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/likert_extrovertism.jpeg

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/likert_neuroticism.jpeg

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/likert_openness.jpeg

Bar charts were produced to compare the difference in mean score for each trait and each Likert type answer. For the traits, the largest difference in means seems to be in Agreeableness, with women scoring higher and Neuroticism, with men scoring higher.

For Likert answers, the largest difference appears in the Moderate category, with men ranking higher. Not surprisingly, to balance out, women rank slightly higher in the other two categories.

Plots:

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/traitmeans.jpeg

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/likertmeans.jpeg

To confirm what was seen in the above plots, statistical tests to performed to see if the means were different for the two populations. First, density plots, with normal overlays, were created to look for normality in the data. All of the features did display close-to-normal behaviour.

Plots:

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/density_agreeableness.jpeg

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/density_conscientiousness.jpeg

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/density_extrovertism.jpeg

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/density_neuroticism.jpeg

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/density_openness.jpeg

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/density_extreme.jpeg

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/density_moderate.jpeg

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/density_neutral.jpeg

To test for equality in variances, F-tests were performed for each feature.

Results:

https://github.com/davidlichacz/Capstone_Project/blob/master/Test Results/ftestresults.txt

While the results of a couple of the F-tests do indicate that there is enough evidence to suggest that the null hypothesis of equal variance is correct, there are other results where this is not the case. As a result, Welch's variant of the t-test was used to test for equality in the means since it is more reliable when the two samples have unequal variances and unequal sample sizes.

Results:

https://github.com/davidlichacz/Capstone_Project/blob/master/Test Results/welchttestresults.txt

The results of the Welch t-test show that there is evidence to suggest that the means in every feature differ between genders. In the case of the personality traits, the most noticeable differences occur in Agreeableness and Neuroticism. In the Likert categories, the biggest difference is seen in Moderate. Both of these results confirm what was seen in the visualization steps.

Step 4: Initial Results

R Script: https://github.com/davidlichacz/Capstone_Project/blob/master/initialresults.R

R Output: https://github.com/davidlichacz/Capstone_Project/blob/master/initialresults.txt

To obtain initial results, five different classification algorithms using an 80% training set were run using all 50 test questions in an attempt to predict gender. (Note: AdaBoost uses the entire dataset rather than training/testing.) The results are summarized below:

Method	Overall Success Rate	Male Success Rate	Female Success Rate
k-NN	62.0%	44.0%	73.1%
Naive Bayes	65.9%	53.0%	73.9%
Logistic Regression	68.4%	46.5%	82.0%
Random Forest	68.0%	40.4%	85.0%
AdaBoost	69.9%	47.7%	83.9%

ROC curves were plotted for comparison:
https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/roc_initial.jpeg

Since the AdaBoost algorithm produced the highest overall success rate as well as the greatest area underneath the ROC curve, this is the model that was chosen to move forward in the analysis.

Step 5: Feature Reduction

R Script: https://github.com/davidlichacz/Capstone_Project/blob/master/featurereduction.R

In an attempt to reduce the number of features, and therefore produce a shorter questionnaire, the list of important features as determined by the AdaBoost algorithm was analyzed. A loop was created that dropped the least important feature at each iteration and then calculated an updated success rate. The results are summarized in the following plot:

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/accuracyrate50.jpeg

In looking at this plot, it was determined that the model with 25 of the most important features (a 50% reduction rate) provided a good balance between accuracy rate and the number of features used. At this point, a new model was fit using the 25 features and the above process was repeated producing the following plot:

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/accuracyrate25.jpeg

In this case, the plot is mostly monotonically increasing with no obvious place for more effective feature reduction. Therefore, the 25-feature model will be used moving forward.

The features that will be used are: A1, A2, A3, A4, A5, A6, A7, A8, A9, C1, C8, C9, E2, E6, E9, N1, N10, N2, N6, N7, N8, O10, O2, O5, O8. Each of the five traits is represented in the reduced list of features with the greatest number coming from Agreeableness and Neuroticism. This is consistent with the results seen from previous exploratory analysis.

Step 6: Model Refinement

R Scripts: https://github.com/davidlichacz/Capstone_Project/blob/master/modelrefinement1.R

https://github.com/davidlichacz/Capstone_Project/blob/master/modelrefinement2.R

To attempt to improve classification accuracy, the Likert answer data was added to the model. This was done in two different ways: First, the original Likert data was added to the 25 important features and a model was produced. However, since this Likert data was calculated using all 50 features, predicted Likert data using the 25 important features was calculated and applied in a second model. The ROC curves were calculated and plotted:

https://github.com/davidlichacz/Capstone_Project/blob/master/Plots/roc_likert.jpeg

As seen, the Likert data produced minimal, if any, increased accuracy and, unfortunately, will be left out of future models.

Also, all models up to this point have seen a large gap in the accuracy in which women are predicted versus men. Since the original dataset consists of roughly 61% women, datasets were tested that used various proportions of men and women: a balanced 50/50 split, 80/20 and 20/80.

The results are summarized in the following confusion matrices:

https://github.com/davidlichacz/Capstone_Project/blob/master/gendersplitmodels.txt

While the highly unbalanced models produce near perfect results, it was suspected that this was the result of overfitting due to the oversampling of the majority class. In fact, when this model was performed on separate training and test sets, the accuracy rates for the minority class were extremely poor.

On the other hand, the balanced model produces a similar overall accuracy rate to those seen previously, but the gap between the two classes is much smaller than previous models. When performed on separate training and test sets, there are some signs of overfitting due to the oversampling of men, but there is still a greater balance between genders than previous models. Since the results of this study would be applied to a general population, this balanced approach will be incorporated into the final model.

https://github.com/davidlichacz/Capstone_Project/blob/master/gendersplitmodelstest.txt

Furthermore, when the importance of the variables in these models is analyzed, very little difference is seen in the variables of the imbalanced models, while the balanced model returns more interesting results.

Step 7: Final Model

R Script: https://github.com/davidlichacz/Capstone_Project/blob/master/finalmodel.R

To prepare the final model, a final balanced dataset was created that used all of the Female instances in the original dataset and oversampled the Male instances. This dataset consists of 23992 instances with a gender column and 25 predictor variables.

The train function from the caret packages was used to optimize the parameters for the AdaBoost algorithm. The suggested parameters were: mfinal=150, maxdepth=3 & coeflearn=Zhu.

The final model was fit using the optimized parameters and produced the following confusion matrix:

https://github.com/davidlichacz/Capstone_Project/blob/master/finalresults.txt

Conclusion and Next Steps

The final model produced an overall accuracy rate of 66.6% and a 95% confidence interval of (0.66, 0.6719) with a Male accuracy rate of 66.65% and a Female accuracy rate of 66.55%. It is expected that the performance on an independent dataset might be somewhat lower for the Male population due to oversampling in the dataset.

The accuracy rates found in the model were lower than were hoped for and would likely not be sufficient to be used in research or business decisions. The reduction in features from 50 to 25 is also less than expected as 25 questions is a significant addition to a questionnaire and could contribute to survey fatigue.

There are some positives to be taken from the results, however. The accuracy rates are significantly higher than the no-information rate, suggesting that five-factor personality results can play a role in predicting gender.

Furthermore, since any tool to create anonymous surveys should be designed for the general public, the similarity in accuracy in Males and Females is a great improvement over the initial results. The results of McNemar's test suggest that the row and column frequencies of the confusion matrix are equal.

The following steps are proposed to continue the study:

- Further investigation into the Likert response patterns: It is still believed that the Likert responses could play a role in the classification, but perhaps a different metric is needed to quantify them.

- Further feature reduction: Due to time constraints, only the iteration through important variables was considered, but perhaps other combinations of features could yield more positive results.

- Investigation of a larger dataset: Perhaps a larger dataset is needed to capture all of the nuances of a complex subject such as human personality.

- Consideration of other types of questions: While the five-factor model has provided a basis for gender classification, we may have reached its limitations and need to supplement it with other data.

- Other classifications: Once a sufficient model for gender is found, other demographic categories, such as age and race, can be considered.

References

Costa, P.T. Jr.; Terracciano, A.; McCrae, R.R. (2001). "Gender Differences in Personality Traits Across Cultures: Robust and Surprising Findings". *Journal of Personality and Social Psychology 81*(2): 322–331

Goldberg, L.R., (1990). "An Alternative "Description of Personality": The Big-Five Factor Structure". *Journal of Personality and Social Psychology, 59* (6), 1216-1229

Joinson, Adam, (1999). "Social Desirability, Anonymity and Internet-Based Questionnaires". *Behavior Research Methods, Instruments & Computers, 31* (3), 433-438

Ong, Anthony D.; Weiss, David J. (2006). "The Impact of Anonymity on Responses to Sensitive Questions". *Journal of Applied Social Psychology 30*(8): 1691–1708