**Personalities and Their Hobbies**

**Emilio Ramos Monzalvo**

**Syracuse University**

**Abstract**

The Kaggle competition brings into question how to separate different Personality Groups using different interests. The 270 interests have been encoded to prevent the user to make assumptions based on previous knowledge. Therefore, the problem is honest and has no types of misdirection. The 270 interests can only have a positive response or no response which makes the dataset extremely sparse. Preprocessing the dataset into a computer friendly manner was the first task in order to plug in the values into a Machine Learning algorithm. On the other hand, preprocessing the dataset too thoroughly can bring in overtraining. Especially since there were only 6340 individuals who participated in the study. On a survey scale 6000 people is a substantial number of participants, but when it comes to Machine Learning, the more data the better results. The main tool of choice was the Random Forest and Decision Tree. The preprocessed dataset was used in the algorithms and yielded a 92% accuracy with Supervised Learning. The results also returned a visual representation of the most thought-provoking interests with the help of a graphical representation of the Decision Tree. The above analysis can not only predict where an individual falls into out of the 4 groups, but can also bring insight of what interests an individual might want to explore after knowing the type of personality they have.

**Introduction**

In order to understand individual’s desires, the Kaggle dataset called “Clustering Categorical Peoples Interests”, has been posted online for Data Scientists to cluster people’s hobbies by their personalities. The post wants to map personalities to hobbies as a way to give recommendations to people after completing a personality test. In a simple example, they would like to know if Group ‘C’ has a high likeliness to be involved in Interest5, but Interest115 is not very likely to part of their routine.

The dataset is sparse since a questionnaire was given to 6340 individuals asking them if they were interested in 217 hobbies. The responses to each question are depicted by 1 or nothing in the dataset where 1 represents an interest in the hobby and nothing represents no interest. Each hobby has been renamed as Interest(N), where N is a number in between 1 and 217. This encoding was done in order to prevent the researcher from making biased understanding from the data.

The main challenge with this dataset is finding the columns that well represent the interests of the groups. As previously mentioned, the dataset contains 217 interests which corresponds to only 6340 individual responses. The algorithms will have to be carefully picked in order to avoid overfitting. The data will also have to be preprocessed extensively in order to not leave out any important data, but still dropping the irrelevant columns that have no contribution to the analysis. This is due to the fact that the dataset is extremely sparse, so many measures have to be taken before feeding the data into different Machine Learning algorithms. The computation time of 270 columns will overload the computer and lead it to restarting the kernel.

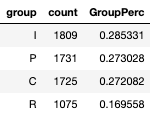
**Preprocessing Data**

The Kaggle competition claims that the dataset has been preprocessed fully before being posted online for simplicity. On the other hand, there is still 217 columns for 6340 individuals; therefore, the first step will be to take out columns with small number of positive responses. Out of the 217 interests, 177 of them had more than ten positive responses. The 40 columns will be treated as outliers and removed from the study. If only ten people responded positively to the interest, it will not lead to a true representation of a personality group. Then there are two  
 columns which have more than 5990 positive feedback from the 6340 individuals which are also outliers and consequently will be taken out of the study as well. After this dropping these 42 columns, there are 175 interests left.

Now while pre-analyzing, the data, there were a couple of cases found that instead of 1, the dataset actually had a 2 which could make our predictive algorithms not know how to compute them. The 2’s were replaced by 1 instead for a common value for True responses by individuals.

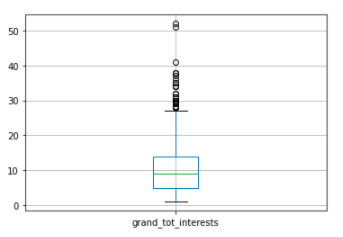
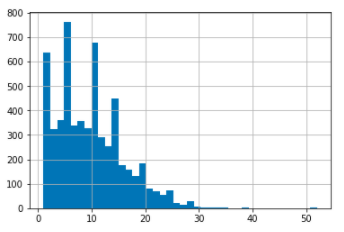
The next step to cleaning up the dataset is getting to find interests that are well divided among the four categories. To perform this analysis, we first need to normalize the percentages according to the population of the categories in the dataset. That means, that we need to balance the percentages using the total population of each group in the dataset. Table 1 below depicts the imbalance of the dataset. Group R is the smallest group and can cause problems when we analyze the whole dataset without normalizing the data.

Table 1: Number of Individuals in Each Category



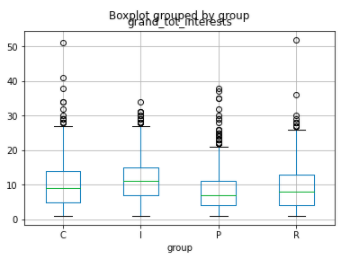
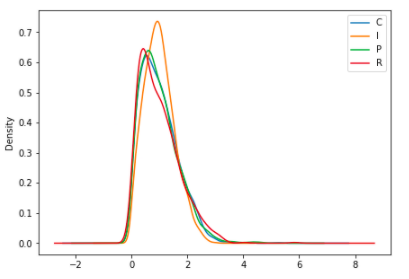
With the normalized percentages, there was about 51 columns where there was no significant difference between all four categories of individuals that like that interest. For example, for interest6, we found that the normalized percentages are 23.1%, 26.5%, 23.8%, and 26.6% for group C, I, P, and R respectively. Therefore, there is no significant difference among the groups for this particular interest. After this analysis we are left with 124 interests.

The next step would be to get of the anomalous rows or individuals. To be anomalous the individual needs to have either a big amount of positive feedback or a very small amount relative to the whole population. In Figure 1, the distribution of the number of interests per individual is displayed after filtering out the individuals with zero number of interests. The idea that that an individual can have zero interests would only beneficial if one group dominated having no interests, but that is not the case. Therefore leaving 5788 remaining individuals in the study.

Figure 1: Distribution of Number of Interests by Whole Populations

According to the distribution displayed above, any individual with more than 28 interests is considered an anomaly. The down side of this approach is that the individual might be an anomaly in the whole population, but might be representative of a particular personality or group. Conversely, in Figure 2, the distribution per group is displayed.

Figure 2: Distribution of Number of Interests by Group



Even by looking at this distribution, the outliers are still above 28 number of interests. Consequently, the individuals with more than 28 interests were removed from the study leaving 5737 remaining participants.

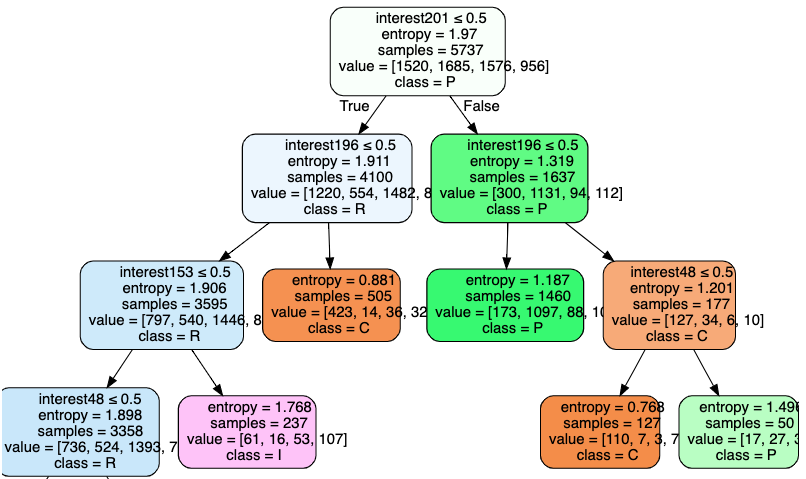
**Random Tree best Entropy Based Splits**

In order to continue finding which interests are the most unique between categories, we can train a random tree model and use the computed edges in the tree model to find out which parameters cause the biggest reduction in entropy when split in half. To perform the analysis, a Random Forest model was trained using a balanced dataset since an imbalanced one can cause overfitting. There was cross validation and tuning done using the Grid Search CV function in Sci-Kit Learn. The parameters that were tuned includes the number of estimators/trees of the forest, and the max depth while the criterion was left as entropy. In the end, the best model was the random forest with depth equal to 20 and number of trees equal to 50. Then using the built-in method in Sci-Kit Learn to get the parameter importance, the top 30 parameters were selected.

**Visualizing Using Decision Tree**

Using the 30 most important features, there was a Decision Tree Classifier trained using Grid Search CV tuning the max depth, max features and the max leaves, but keeping the numbers small in order to be able to visualize the decision tree. The Decision Tree with the best results had a depth of 20, 15 features and 10 leaf nodes. The prediction accuracy was 56.6%, but since the intent of this analysis is to understand the biggest differences between the types of personalities, we can view the top splits in the Decision Tree in order to get a visual understanding of the differences.

Figure 3: Decision Tree Top Splits

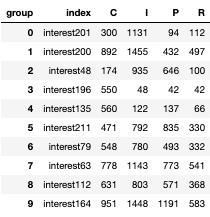
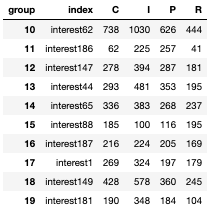
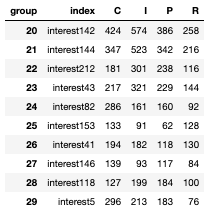


The Decision Tree in Figure 3 gives a good idea on the most divisive parameters that can lead us to an understanding of the personalities. Interest 201 is the most divisive of all features and clearly separates Personality R and Personality P since the criterion of the decision tree is set to entropy, and Personality P and Personality R are only seen in their respective branches of the tree after the first split.

**Conclusion**

In Figure 4, the top 30 controversial parameters are displayed with their corresponding count per personality group. However, competition requirements were looking for a set of parameters that can truly represent these 4 groups. In the approach taken, the full set of interests were preprocessed using simple arithmetic properties and machine learning to get rid of interests with no argument. This includes finding interests that are well represented among all groups and interests that have a very limited amount of positive responses. After the first round of filters, a machine learning approach was taken. A Random Forest Approach was used to find the top 30 contributing parameters and yielded a 92% accuracy. The main idea of using this approach is to use Entropy to find the interests that are best in splitting up the group. For example, interest 196 shows a clear influenced for Group C compared to the rest of the groups. These results were also used in a Decision Tree to be able to visually see the splits. All in all, the top 30 parameters should show a clear split between the clusters.

Figure 4: Count of Individual Interests per Group

**References**

1. “Young People Survey.” *Kaggle*, www.kaggle.com/miroslavsabo/young-people-survey.