Predicting Young People

Evan Feiereisel, Robert Henry Quan
Department of Computer Science, Texas A&M University, College Station, Texas 77843-4242, USA
(Dated: December 11, 2017)

I. ABSTRACT

Through the use of SVMs and Deep Neural Networks, we classified and predicted preferences, interests, habits, opinions, and fears of young people. By using the mean squared error and the accuracy rating as our evaluation, we computed the ideal parameters for our models using a 3 k-fold cross validation algorithm. Our Deep Neural Network was trained to perform regression, using no activation functino, and having multiple outputs on the last layer. Our SVMs were trained to perform multi-class classification, building classifiers for each pair of classes, and using a decision function to combine the results.

II. DATA

For our data, we decided to use the Kaggle data set Young People Survey which explores the preferences, habits, interests, opinions and fears of young people. The survey was implemented by the students of the Statistics class at FSEV UK, located in Bratislava, Slovakia. The research questions consist of a numerical score in the following groups: music preferences, movies preferences, hobbies interests, phobias, health habits, personality traits, views on life opinions, spending habits and general questions on demographics of the takers of the survey. The data file consists of 1010 rows and 150 columns (139 integer responses and 11 categorical). The data set was not perfect and did contain missing entries for some of the questions. The survey was presented to the participants in both electronic and written form. Also, the original survey was in Slovak language which was later translated into English. All of the participants were of Slovakian nationality, aged between 15-30.

III. OBJECTIVE

Our mission in analyzing this data set, is to find interesting correlations between the different groups of preferences and based on this we can further analyze the best indicators of types to predict other groups types. For example, if a response has a strong disliking for Country music, there is a high probability that the individual will have have a strong disliking for western movies.

IV. DATA PREPROCESSING

To begin our processing, we import the .csv files with the python pandas library and then separated the columns into different groups. The groups we synthesized were: music, movies, hobbies, demographic, spending habits, phobias and personality. Each group of columns had 1010 rows which signifies that 1010 individual responses the survey. Some of the values for these groups were missing and filled with a Nan value. With pandas, we could easily analyze these missing values, and based on the repose, we did fill in those values with a neutral response. In the data, there were a small number of categorical features. We decided to disregard the columns that were categorical for our project.

V. IMPLEMENTATION

For our regression models we used MLPRegressor from the sklearn library, which implements a multi-layer perceptron that is trained on samples through backpropagation with no activation function in the output layer, and uses square error as the loss function. The reason why we decided to use the MLPRegressor is due to the fact that the network can be trained on multi-output regression that are represented as the groups of data which we are interested in finding a correlation with. For example, from the music input group, we can create models that can predict which movies, hobbies, demographic, spending habits, phobias and personality traits the individual might have with a certain accuracy.

For our model, we set the number of hidden layers to be 150, an activation function that defaulted to rectified linear unit function, a solver that was set to adam and an alpha value of 0.001. The solver adam refers to a stochastic gradient-based optimizer proposed by Kingma, Diederik, and Jimmy Ba. We chose adam because it works better with large datasets and performs better for the validation score. The alpha parameter is the L2 penalty (regularization term) parameter to prevent overfitting our data.

After some research, we found that the sklearn kit has a predefined function cross_val_score which takes and arguments an estimator, multi input-Vector, multi output-Vector, and a scoring function. The function also defaults to a 3-fold cross-validation generator to better

average the scoring function. For our regressino model, we chose a scoring function of negative mean squared error.

We used a SVM for our classification models. Since our output data was a score from 1-5, we used a multi-class SVM, training models for each pair of classes, and then combining all of them with a decision function. We did this for entire groups to classify one attribute, as well as individual attributes to classify individual attributes. We also used the cross_val_score from sklearn to perform our training.

VI. EVALUATION

In evaluating our model, we used 3 fold cross validation with a scoring function of mean squared error. The reason why we use a 3 fold cross validation, is to take the mean of the mean squared error from the output and judge which groups of data will better predict other groups of data.

This allowed us to create models for all combination of preferences. For example, we can now predict movie preferences on music preferences, or spending habits on demographics.

We used the same template, of 3 fold cross validation, but with a scoring function o0f accuracy for our SVM classifiers.This allowed us to create models for group-attribute data, and attribute-attribute data.

VII. RESULTS

Below we see results of our reression and classification for different levels of granularity.

A. Group-Group Regression

Performing multi-dimensional regression with a NN, we predicted every group based on every other group. Below are the mean squared error for each of the pairs.

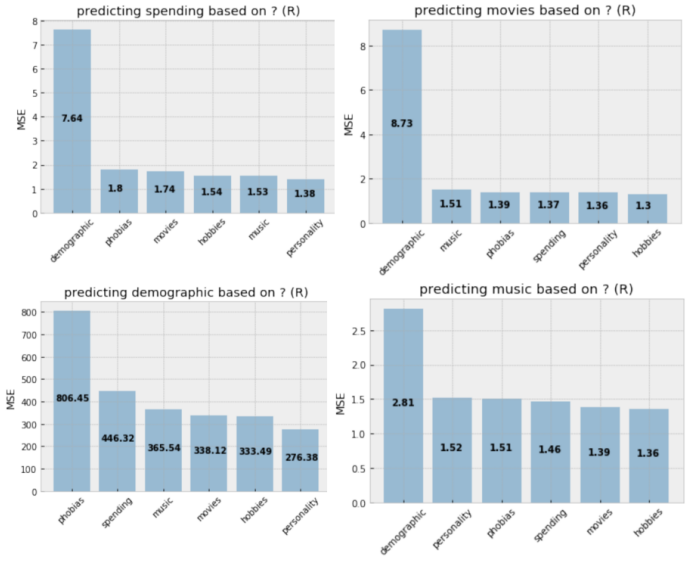

FIG. 1: Mean Squared Error rating per Group-Group Regression

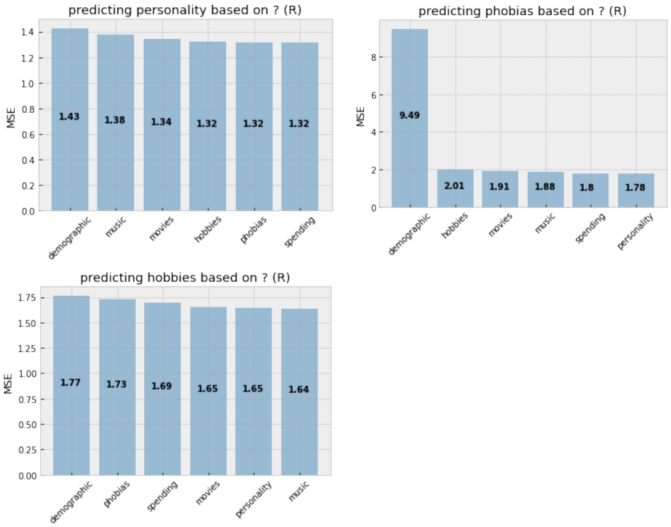

FIG. 2: Mean Squared Error rating per Group-Group Regression

B. Group-Attribute Classification

From the Group-Group Regression, we chose the pairs with the least mean squared error, to perform Group-Attribute Classification. Here we took a group, group pair, and split the group being predicted into each of it's features, and predicted classification on each, to see which we could predict with highest accuracy. Below are graphs of a group predicting all features of another group, with accuracy as the scoring function.

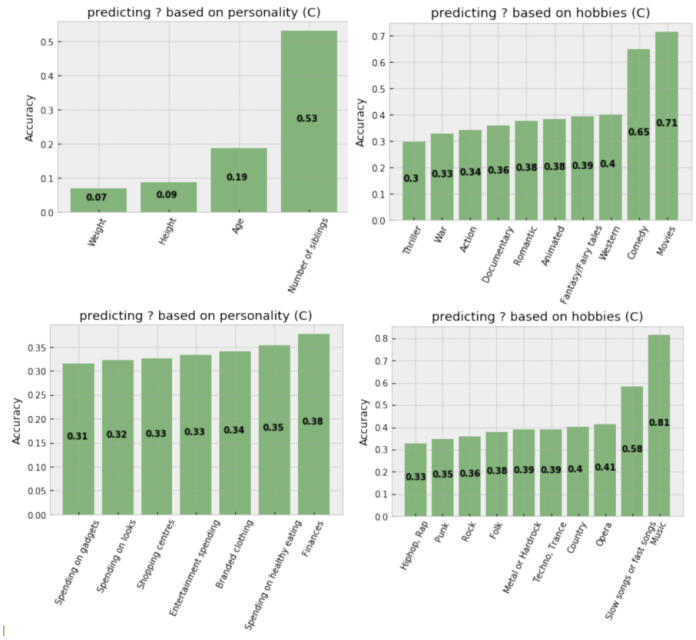

FIG. 3: Accuracy rating per Group-Attribute Classification

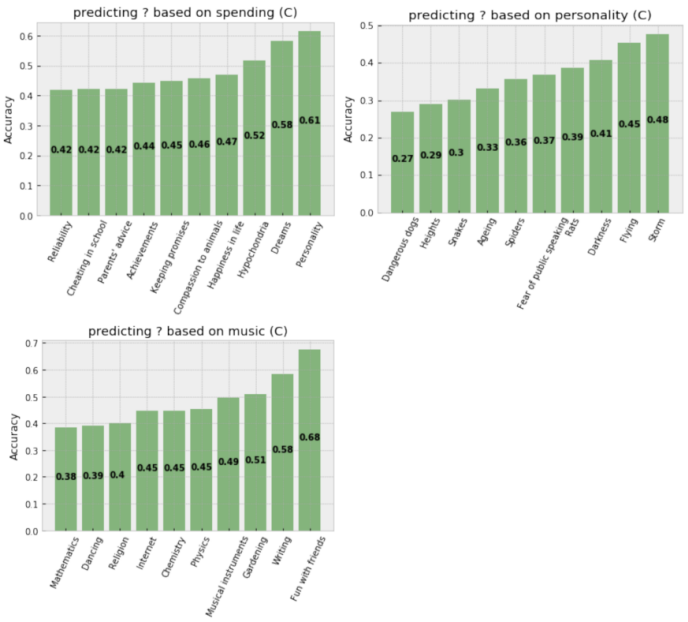

FIG. 4: Accuracy rating per Group-Attribute Classification

C. Attribute-Attribute Classification

From the Group-Attribute Classification. For each Group, we found the best attribute it could classify, and then went into the predicting Group and split it up to perform attribute attribute classification, for each attribute in the predicting group, and for the best attribute found from the group-attribute classification. Below are a few of the results.

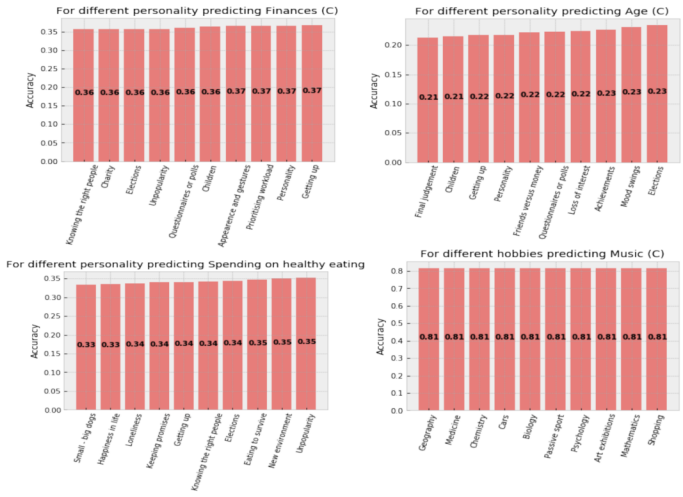

FIG. 5: Accuracy rating per Attribute-Attribute Classification

VIII. DISCUSSION

From our results, we found a couple interesting connections. We found that both spending habits, and personality were best predictors of each other. Within that, we found that we could best classify someone's view of their own personality based on their spending habits, with 61 percent.

We also found that we could best predict demographics based on personality. Within this pair, we found with 53 percent, we could classify the number of siblings based on personality.

Another interesting result, was that we could best predict music based on hobbies. Within this pair, we could, with 58 percent accuracy, classify if you like fast or slow music based on hobbies. On an attribute to attribute level, we found an equal classifying rate for all hobbies for classifying how much you like music. Indicating that no individual hobby has much classifying power for that attribute. For this question, and a couple others, it seemed like most people answered in a similar manner, and the answer had little to no correlation with other questions.

IX. CONCLUSION

Through the use of Deep Neural Networks, and SVMs, we explored the data of young people preferences in music, movies, spending habits, etc. Using regression we tried to predict a whole preference based on another. Using classification we increased the granularity, and looked into individual attributes. We found some interesting results, such as highest predicting power between spending

habits, and personality. We also found that some questions were very common in their answers, which had the classifiers classifying more on the majority of answers, rather than using other data to learn mappings.

[1] Sleziak, P. - Sabo, M.: Gender differences in the prevalence of specific phobias. Forum Statisticum Slovacum. 2014, Vol. 10, No. 6.
[2] Sabo, Miroslav. Multivariate Statistical Methods with Applications. Diss. Slovak University of Technology in Bratislava, 2014.