Does Personality Predict Achievement?
Project 3: McNulty for Metis SF Spring 2018.
Gray Davidson

PROJECT DESIGN:
The goal of this project was to explore a complex psychometric dataset with machine learning, compared many algorithms and developed the best possible supervised model to predict college attendance from personality and grit questionnaire data. The data were downloaded, engineered, cleaned, trimmed and fed into six different classification algorithms. These were compared on the basis of AUC and either accuracy or F1 score depending on the scope of class imbalance. Ancillary analyses of Feature Importance, PCA and Logistic Regression Coefficients sought to relate the overall findings to the individual features of the data set with mixed success.

Historically personality was thought to be fixed in an individual’s genes but more modern research has shown that the five factors continue to evolve over a lifetime and are subject to some degree of intentional control. Grit likewise is not fixed so the predictions from this model become an invitation to work on ourselves and of course environmental factors also play a big role in individual achievement. Thus future directions for data collection in this field would be (and have been) to examine socio-economic status, location, etc. as a part of the dataset and to examine the persistence of personality factors through time.

DATA:
The data were drawn from:
http://openpsychometrics.org/_rawdata/duckworth-grit-scale-data.zip.

These data are the combined answers by 4000+ subjects to Big 5 Personality, Grit and Demographic questionnaires. For demographic data subjects answered questions such as “What is your sexual orientation?” by rating 1=Heterosexual, 2=Bisexual, 3=Homosexual, 4=Asexual or 5=Other. The five personality factors and grit were assessed on a Likert scale (Subjects would rate how much they agree or disagree with a statement such as “I am the life of the party” on a scale from 1 to 5). Ten such statements were averaged for each of the personality axes and twelve were averaged for Grit.

To prepare the data for modeling individual question scores were aggregated into overall scores for each personality axis. Categorical data were divided by column such that each of the twelve religion choices had its own column and a given subject would score 0 in all columns except the one they had selected where a 1 would appear. I also shifted the questionnaire data to have a zero-mean. Finally I spent a fair amount of time detecting and accounting for outliers, missing values and impossible values (such as a 0 answer for a question which had only natural number answers).

I selected education as the classification variable because this question interested me — could I predict a person's educational achievement from their personality matrix? Education was reported in four levels: 1 = "Some high-school" 2 = "High-school graduate" 3 = "College Degrees" and 4 = "Graduate degree." I ran two separate classifications, one to distinguish between high-school and college, and the other to distinguish between college and graduate school. In the former case the two classes were split fairly evenly in the data but in the latter case there was a 6:1 ratio in class balance. To counteract this I oversampled the "Graduate Degree" class, grid searching across many different ratios and concluding that the best F1 scores were to be had at a 1:1 ratio.

ALGORITHMS:
After coding and tuning six different classification algorithms:

Naive Bayes:
KNN: (Grid Searched in log space then linespace for ideal k)
Logistic Regression:
Support Vector Classifier: (Grid searched for C and four kernels)
Decision Tree: (Grid searched for several parameters)
Random Forest: (Grid searched for several parameters)

The results were summarized via a visualized receiver operating characteristic (ROC) curve where chance was displayed along with all six models. Accuracies ranged from .55 to .82. The ROC curve shows how various models relate to one another by capturing true-positive examples and avoiding false positives. A valid metric for comparing these algorithms is the area under the curve (AUC) which provides an aggregate measure of performance across all possible classification thresholds.
Random forests (1500 estimators, 40 features maximum, 2 samples/leaf minimum) resulted in the highest AUC and also clocked in with the highest accuracy of any model, correctly predicting a participant's education level 82% of the time. With so much subtlety to the construction of the dataset (eleven items with slight variations coding for the same concepts) it is not surprising that complex methods were required to find the best result. In a simpler space (say the features were restricted to the five aggregate factors and grit) a simpler model such as KNN may well have done a good enough job.

TOOLS:

Important Features Analysis from the Random Forest model highlights Big5 and Grit as opposed to demographic data. Primary Component Analysis (PCA) run on the un-modeled data demonstrated similar results, showing that the strongest component in the data was the Extraversion/Grit/Conscientiousness while the second was Neuroticism. PCA components were examined separately and were not included in the dataset for analysis. Logistic Regression Coefficients showed no clear result and were discarded.

As mentioned above six classification algorithms were used, (Naive Bayes, KNN, LR, SVC, DT, RF). Pandas was used extensively to keep data orderly and to prepare it for the estimators. Matplotlib was used to create some visualizations and the remainder were constructed in Google Slides or were used under Creative Commons.