Predicting Entrepreneurship

Matthew Keiser

The United States Census Bureau performs a survey called the American Community Survey every year to track demographic and social trends across the United States. Among the features tracked by this survey is “Class of Worker” which describes how a person works - private, public, non-profit, self-employed, farm, unemployed. It would be valuable to be able to estimate the number of potential or actual entrepreneurs of a population using demographic data. With that in mind, this project attempts to do exactly this.

The data set contains 3,132,795 observations and 283 features. Many of the features are allocation flags and other non-informational fields that were removed. All entries for persons less than 18 years old were also removed to target adults. After these manipulations, the dataset is ready to be analyzed. The “Class of Worker” feature was used to create a categorical field that splits the entries between entrepreneurs and non-entrepreneurs. The features were split between categorical features and continuous value features.

The dataset was split into 10 even subsets to allow for efficient operation of the modeling and prediction activities, with one of the subsets set aside for validation. To create the estimator functions, each subset was appended with dummy variables of all of the categorical features, and a Random Forest Classifier was fit to each subset using grid search cross validation to determine the best number of trees to use for each classifier. Finally, each classifier function was saved as a pickle file, to avoid fitting the model each time the analysis is done.

To recombine predictions from the classifiers, a vote is taken (using the scipy.stats.mode function) using the outputs from the 9 classifiers. A prediction was made on the test data, and an accuracy score was calculated. The final accuracy score was 97.3%.

Challenges in this analysis are directly related to the size of the dataset. The dataset splitting was required to run the analysis due to memory constraints. Training the models on the subsets took approximately 20 minutes per model, making re-training take a long time. By saving the subsets as CSV files to be read in on demand, memory issues were avoided. By saving the classifiers to pickles files to be read in on demand, the classifiers did not have to be re-trained each time the model is run.

An expansion of the project would be to perform a multinomial classification by predicting each Class of Worker, as opposed to only Self-Employment. Due to the success of the model on this narrow classification, it is likely the extended analysis would generate an accurate prediction.