**Race/BPL Prediction (1/22 - 6/28)**

Michael Cao

**Objective**

* Produce predicted probabilities of race and foreign/native-born status for loan card records based on the 1940 US census.

**Source Files**

* Projects(S:) / CR4230 / Projects / Project Collaboration / Michael’s Files
  + 1940\_census\_cleaned.parquet
    - Also available in .rds and .csv
    - Raw files available in Projects(S:)/CR4230/1940\_2.0
  + card\_data\_formichael\_2024\_04\_24.rds
  + Data Recoding New (folder)
  + Requested Tables 043024 (folder)
    - Most recent version: includes 5- and 2-race scenarios
  + PPL\_HHD.Rmd [1]
    - This script manipulates the partitioned raw people & household data and the .yml (data dictionary) file to produce 1940\_census\_cleaned.parquet for model usage.
  + Race Prediction.Rmd [2]
    - This is the main code that includes the logic of the BISFG algorithm (with modifications) that produces the predicted race probabilities of the loan card records.
    - Further cleaning/condensing of the census file is also included to enhance efficiency (though the parquet file produced from PPL\_HHD.Rmd is not affected).
    - This file also includes the backtesting of the algorithm, which is the content of Table 7 (in requested tables).
  + Table Requests.Rmd [3]
    - This code file takes in the census data and loan card predictions and produces the requested tables.
  + BPL Prediction.Rmd [4]
    - This code provides similar outputs as the second .Rmd file but for predicting native/foreign-born status.

**Methods**

* The prediction algorithm uses Bayesian Improved Surname First-Name Geocoding (BISFG) with additional relaxation steps. In particular, the algorithm only continues if the previous step returns NA or if the predicted probabilities are equal in the top two categories. The order of the steps is as follows:
  + Run BISFG with first name (of head of the household), last name, and state;
  + If first name (of head of the household) only contains 1 letter, or if the continuing condition occurs, then BISFG is run again with the spouse’s first name, last name, and state;
  + If the continuing condition still occurs, or if the spouse’s name only contains 1 letter or if it doesn’t exist, then BISG (using only the last name) is run;
  + If the continuing condition still occurs, the prediction is left as it is (even if the predicted output is NA).
* The backtesting method involves two metrics - an accuracy confidence interval and an AUC score (or C-statistic), computed using sample census data from each state and overall.

**Results**

* Of the 39309 loan card records in the available states, the algorithm is able to output a prediction for all but 866 records (almost all of them due to uncommon last names), or 97.80% of all records.
* An accuracy of around 80.5% is achieved for race prediction with five categories. This accuracy improves to approximately 83.5% when reduced to two racial categories.
* An accuracy of around 81% is achieved in native-foreign prediction.
  + [ADDITIONAL DETAILS/OBSERVATIONS TO BE ADDED…]

**Additional Notes**

* The following paper is referenced in creating the algorithm:
  + *Voicu, I. (2018). Using First Name Information to Improve Race and Ethnicity Classification. Statistics and Public Policy, 5(1), 1–13.* [*https://doi.org/10.1080/2330443X.2018.1427012*](https://doi.org/10.1080/2330443X.2018.1427012)