Instructions for Using Race Imputation Model from "Explaining Racial Disparities in Personal Bankruptcy Outcomes"

Bronson Argyle[*] **Sasha Indarte**[†] **Ben Iverson**[‡] **Christopher Palmer**[§]

April 4, 2024

1 Conditions for Use

In any work using our race imputation model, or any of the code or data provided along with the model files, cite our paper as the source.

> Argyle, Bronson, Sasha Indarte, Benjamin Iverson, and Christopher Palmer. "Explaining racial disparities in personal bankruptcy outcomes." Working paper, 2023.

The race imputation code described below, along with the needed Census data, is available as an 872 MB zip file at `https://mit.edu/cjpalmer/www/AIIP-Race-Imputation-Code.zip`.

2 List of Files Needed to Run Imputation Model

Important: our model was written using the 2.12.0 versions of Python's Keras and TensorFlow libraries. Some of the newer versions of these packages are not compatible with our syntax and .pkl files and will generate error messages.

- Model files (located in model_files folder

  `model_block.pkl`: model that uses block-level race composition data `model_tract.pkl`: model that uses tract-level race composition data `model_zip.pkl`: model that uses

[*]Brigham Young University. Email: **bsa@byu.edu**
[†]The Wharton School of the University of Pennsylvania. Email: **aindarte@wharton.upenn.edu**
[‡]Brigham Young University. Email: **ben_iverson@byu.edu**
[§]Massachusetts Institute of Technology and NBER. Email: **cjpalmer@mit.edu**

ZIP-level race composition data `model_county.pkl`: model that uses county-level race composition data `char_vocab.pkl`: encoding of bigrams `word_vocab.pkl`: encoding of words

- `AIIP_race_imputaton_MWE.py`: this file illustrates how to use our model files to impute race. It applies the imputation to four names in "sample_names/sample_name_file.csv"
- The folder data_census_race_comp contains 8 csv files. Each records race composition for the 5 race categories used in the imputation (Asian, Black, Hispanic, Other, and White). The files specifically report the % of the population indicating that they belong to any of these categories. Other includes people identifying as Native American or selecting "other". The files are available for the 4 geographies noted above and for the years 2010 and 2020. The underlying data are the full count Census data from 2010 and 2020. Geocodes for both files are 2020 Census geocodes. Note that if your application uses data from the early 2000s or earlier, you may want to calculate race composition using earlier census data.
- `AIIP_race_imputation_L2.py`: this file contains the code used to estimate our model. Running it requires the L2 data (or similar data containing names and self-reported race). It may be useful for understanding how the model was estimated.

3 Formatting Input Data

How to Format Input Data. Our race imputation model requires two types of variables in order to impute race: names and geographic identifiers. To achieve the best possible model performance, names must be formatted as follows:

Lastname Firstname Middlename

The model uses capitalization to identify the beginning of words. If middle name is unknown or unavailable, we recommend using "Lastname Firstname". If the middle initial is available, it can be included in lieu of "Middlename".

Our model can be implemented using one of four geographic identifiers. From most to least granular, they are: census block, census tract, ZIP code, and county. Specifically, the model uses the local race composition in the individual's residential area, along with their name, to predict

their race. The provided race composition data uses geocodes from the 2020 Census (for both the 2010 and 2020 race composition datasets).