Predicting Race and Ethnicity From the Sequence of Characters in a Name*

Rajashekar Chintalapati† Suriyan Laohaprapanon‡ Gaurav Sood§

June 11, 2023

Abstract

To answer questions about racial inequality, we often need a way to infer race and ethnicity from a name. Until now, the bulk of the focus has been on optimally exploiting the last names list provided by the Census Bureau. But first names contain more information, especially for African Americans. To estimate the relationship between full names and race, we exploit Florida Voter Registration Data which includes self-reported race and ethnicity. We model the relationship between characters in a name and race and ethnicity using various techniques. We find that a model using Long Short-Term Memory works best with out-of-sample (OOS) precision and recall of .80 and .81 respectively. The equivalent numbers for the last name only model are .63 and .65. To illustrate the use of this method, we apply our method to campaign finance data to estimate the share of donations made by people of various racial groups and coverage of various races and ethnicities in the news.

*Data and scripts behind the analysis presented here can be downloaded from http://github.com/appeler/ethnicolr_v2. This paper is a new version of Sood and Laohaprapanon (2018).

†Rajashekar can be reached at rajshekar.ch@gmail.com
‡Suriyan can be reached at: suriyant@gmail.com
§Gaurav can be reached at gsood07@gmail.com

How often are people of different races and ethnicities covered in the news? What percentage of campaign contributions come from African Americans? Are there racial gaps in healthcare delivery? Do minorities face discrimination in borrowing? To answer such questions, we often need a way to infer race and ethnicity from names. Given the important questions at stake, a number of researchers have worked on inferring race from names (see, e.g., Ambekar et al. 2009; Fiscella and Fremont 2006; Imai and Khanna 2016; Rosenman, Olivella and Imai 2022). We contribute to the substantial literature.

Inferring Race and Ethnicity from Names

Approaches

Researchers have used a variety of approaches to infer race and ethnicity from names. Some researchers have taken advantage of the list of popular last names provided by the Census Bureau (see, e.g., Fiscella and Fremont 2006). The approach suffers from two weaknesses—a biased small set of popular names, and a lack of first names. The information in the first name is especially vital for African Americans, whose last names are hard to distinguish from non-Hispanic whites, and whose first names tend to be distinctive (Bertrand and Mullainathan 2004). Others have harvested Wikipedia data on names and their national origins and used crude features of names to build a national origins classifier (Ambekar et al. 2009). The Wikipedia data is small and suffers from a strong bias toward names of popular people. Still others have used private data and homophily to predict national origin (Ye et al. 2017). The limitation here is that the data are private. Still others like us use voter registration data (Sood and Laohaprapanon 2018; Parasurama 2021). The voter registration data also has its own tradeoffs.[1] There is labeled data from only a few states, and not everyone is registered

[1] There is a parallel literature that combines voter registration data with census data to infer race where we have the name and location of a person (Imai and Khanna 2016; Kotova N.d., see for e.g.,).

to vote.

Estimand

We predict the modal race and ethnicity of people with a particular name. In the SI, we show results from two more models. The first model predicts the probability distribution. The second model predicts all the most popular racial categories that make up at least 90% of the people with the name.

Measurement of Race and Ethnicity

Florida Voter Registration data provide self-reported race and ethnicity of people. Even though race and ethnicity are self-reported, the limitations of the reporting instrument mean that the quality of the data is debatable. For one, Florida Voter Registration data treats race and ethnicity as one dimension with Hispanic treated as one category. Second, the instrument only allows crude categorization. For instance, Indian Americans are grouped under Asian and Pacific Islanders.

Why Model?

If you picked a random individual with last name Smith from the US in 2010 and asked us to guess this person's race (measured as crudely as by the census), the best guess would be based on what is available from the aggregated Census file. It is the Bayes optimal solution. So what good are last name only predictive models for? A few things. If you want to impute ethnicity at a more granular level, guess the race of people in different years (than when the census was conducted if some assumptions hold), guess the race of people in different countries (again if some assumptions hold), when names are slightly different (again with some assumptions), etc.

Data

We exploit Florida Voting Registration data for the year 2022 (Sood 2017). The Florida Voting Registration has information on nearly 15M voters along with their race. Given that we have very few people who identify as multi-racial and Native American, we condense them into Other. Our final dataset only has five categories: Asian/Pacific Islander, Hispanic, Non-Hispanic Blacks, Non-Hispanic Whites, and Other (see Table 1).

Table 1: Registered Voters in Florida by Race.

race	n
asian	253,808
hispanic	2,179,106
nh black	1,853,690
nh white	8,757,268

Models

If we had a census of all the names, and name was the only data we had, then the Bayes Optimal Classifier for predicting the race of the modal person with the name is the racial category with the largest probability density when you take the conditional mean. Assuming we do not have the full list of names or that names can have spelling errors, the next most obvious classifier is an edit distance based classifier. Use KNN with an edit distance (or cosine distance) based distance metric. Another way to model the problem is as 'Bag of Characters.' We try this approach with Random Forest and Gradient Boosted models. Lastly, we can model the sequence of characters. To do that, we try both LSTM and Transformer models.

- **LSTM** To learn the association between the sequence of characters in names and race and ethnicity, we estimate an LSTM model (Graves and Schmidhuber 2005; Gers, Schmidhuber and Cummins 1999) on the Florida Voter Registration Data. We estimate the last name model on a title case transformed version of the last name. For the full name model, we concatenate the last name and first name (ignoring the middle name) and again capitalize each word. We split the strings (last name or last name and first name) into two character chunks (bi-chars). For instance, Smith becomes `Sm, mi, it, th`. Next, we remove infrequent bi-chars (occurring less than 3 times in the data) and very frequent bi-chars (occurring in over 30% of the sequences in the data). We use the remaining bi-chars as our vocabulary. In the Florida Voting Registration Data, this leaves us with 1,146 bi-chars in the case of last name only data, and 1,604 bi-chars in the full name data. In the Wikipedia data, the corresponding numbers are 1,946 and 2,260. Next, we pad the sequences so that they are the same size. Finally, we use 20 as the window size for the last name only model and 25 for the full name model.

On this set of sequences, we train a LSTM model using Keras (Chollet et al. 2015) and TensorFlow (Abadi et al. 2016). Before estimating the LSTM model, we embed each of the words onto a 32 length real-valued vector. We then estimate a LSTM with a .2 dropout and .2 recurrent dropout for regularization (Srivastava et al. 2014). The last layer is a dense layer with a softmax activation. Because it is a classification problem, we use log loss as the loss function. And we use ADAM for optimization (Kingma and Ba 2014). We fit the model for 15 epochs with a batch size of 32.

Results

Table 2 presents some metrics that shed light on how well we did with the last name only model in predicting race OOS using the Florida Voter Registration Data. The OOS precision is .79, recall is .81, and f1-score, the harmonic mean of precision and recall, is .78. There is however sizable variation in recall across different racial and ethnic groups. For instance, recall is .95 for whites and just .21 for non-Hispanic blacks.

Table 2: OOS Performance of Various Last Name Models.

	KNN	RF	GB	LSTM	Transformer	Support
Overall						
Hispanic						
NH_Black						
NH_White						
Other						

Compared to the last name only model, we do much better with a full name model. The OOS precision, recall, and f1-score for the full name model is .83, .84, and .83 respectively (see Table 3). The gains are, however, asymmetric. Recall is considerably better for Asians and Non-Hispanic blacks with the full name—.49 and .43 respectively, compared to .41 and .21 respectively. The precision with which we predict non-Hispanic Blacks is also considerably higher—it is 9 points higher for the full name model. Given Asians and Hispanics have more distinctive last names, the improvement in precision in predicting both is smaller—negligible in the case of Asians and 2 points in the case of Hispanics.

Table 3: OOS Performance of Various Full Name Models.

	KNN	RF	GB	LSTM	Transformer	Support
Overall						
Hispanic						
NH_Black						
NH_White						
Other						

Application

To illustrate the utility of the models we have developed here, we impute the race and ethnicity of individual campaign contributors in the 2014 campaign contribution databases (Bonica 2017) using just the Census last name data and the Florida full name model. We then use the inferred race and ethnicity to estimate the proportion of total contributions made by people of different races.

Based on the census last name data, in 2010, about 83.5% of contributions were made by Whites (see Table 4). But the commensurate number calculated using the Florida full name model was nearly 3% more, 86.5%. Moving to blacks, we see a similar story. Based on census last name data, about 10.2% of contributions came from blacks. But based on Florida's full name model, the number is about 2.3% lower, or a 22.2% relative change. The commensurate difference in estimated contributions by Hispanics is about 1% or about 33% relative change. Among Asians, the commensurate difference is about .5% points or about 18% relative change. A similar pattern holds for 2000. We see that Whites contribute less based on Census last name data than Florida full name model.

Table 4: Proportion of Total Amount Donated to Political Campaigns in 2000 and 2010 by People of Different Races/Ethnicities.

	Census		Florida	
race	2000	2010	2000	2010
asian	2.22%	2.74%	2.00%	2.28%
black	11.04%	10.22%	8.93%	7.92%
hispanic	3.24%	4.32%	3.23%	3.31%
white	83.49%	82.71%	85.84%	86.49%

Discussion

We use voter registration data to learn a model between sequences of characters in a name and race and ethnicity. Given poor African Americans tend to have distinctive first names, the biggest advantage of using the full name over last names is in our ability to detect African American names. Many important datasets, such as the campaign contributions dataset, voter registration files of other states, news data, etc., contain information on both first and last names. And we could make better predictions about race and ethnicity by using both first and last names. As we noted above, we also provide a Python package that exposes the models: https://github.com/appeler/ethnicolr/.

Voter registration data has many limitations. First, not everyone is registered to vote, and blacks and Hispanics are especially likely not to be registered to vote (Ansolabehere and Hersh 2011). If the names of those who are not on the voter registration file are systematically different from those who are, our accuracy metrics are likely inflated. Second, the pattern of names in a single state may be different from the names in other states.

References

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard et al. 2016. TensorFlow: A System for Large-Scale Machine Learning. In *OSDI*. Vol. 16 pp. 265–283.

Ambekar, Anurag, Charles Ward, Jahangir Mohammed, Swapna Male and Steven Skiena. 2009. Name-ethnicity classification from open sources. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*. ACM pp. 49–58.

Ansolabehere, Stephen and Eitan Hersh. 2011. "Gender, race, age, and voting: A research note.".

Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American economic review* 94(4):991–1013.

Bonica, Adam. 2017. "Database on ideology, money in politics, and elections (DIME).".

Chollet, François et al. 2015. "Keras.".

Fiscella, Kevin and Allen M Fremont. 2006. "Use of geocoding and surname analysis to estimate race and ethnicity." *Health services research* 41(4p1):1482–1500.

Gers, Felix A, Jürgen Schmidhuber and Fred Cummins. 1999. "Learning to forget: Continual prediction with LSTM.".

Graves, Alex and Jürgen Schmidhuber. 2005. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures." *Neural Networks* 18(5-6):602–610.

Imai, Kosuke and Kabir Khanna. 2016. "Improving ecological inference by predicting individual ethnicity from voter registration records." *Political Analysis* 24(2):263–272.

Kingma, Diederik P and Jimmy Ba. 2014. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* .

Kotova, Nadia. N.d. "A Deep Learning Approach to Predicting Race Using Personal Name and Location (Natural Language Processing)." . Forthcoming.

Parasurama, Prasanna. 2021. "raceBERT–A Transformer-based Model for Predicting Race from Names." *arXiv preprint arXiv:2112.03807* .

Rosenman, Evan TR, Santiago Olivella and Kosuke Imai. 2022. "Race and ethnicity data for first, middle, and last names." *arXiv preprint arXiv:2208.12443* .

Sood, Gaurav. 2017. "Florida Voter Registration Data.".
URL: *https://doi.org/10.7910/DVN/UBIG3F*

Sood, Gaurav and Suriyan Laohaprapanon. 2018. "Predicting race and ethnicity from the sequence of characters in a name." *arXiv preprint arXiv:1805.02109* .

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov. 2014. "Dropout: A simple way to prevent neural networks from overfitting." *The Journal of Machine Learning Research* 15(1):1929–1958.

Ye, Junting, Shuchu Han, Yifan Hu, Baris Coskun, Meizhu Liu, Hong Qin and Steven Skiena. 2017. Nationality classification using name embeddings. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. pp. 1897–1906.

Supporting Information

SI 1 Performance of the KNN Models

Table SI 1.1: Performance of the KNN (K = 5) Cosine Distance model on the test set.

	precision	recall	f1-score	support
hispanic	0.85	0.78	0.81	10775
multi_racial	0.02	0	0	364
native_indian	0.02	0.02	0.02	111
nh_black	0.59	0.42	0.49	4483
nh_white	0.79	0.92	0.85	25614
other	0.18	0.03	0.04	582
weighted_avg	0.76	0.78	0.76	43433

SI 2 Performance of the Random Forest Models

Table SI 2.1: Performance of the Last Name Random Forest model on the test set.

	precision	recall	f1-score	support
asian	0.09	0.05	0.06	6,867
hispanic	0.69	0.66	0.68	38,961
nh_black	0.34	0.17	0.23	13,726
nh_white	0.6	0.73	0.66	60,970
other	0.17	0.15	0.16	14,374
weighted_avg	0.53	0.56	0.54	134,898

Table SI 2.2: Performance of the Full Name Random Forest model trained on 2M records on the test set.*

	precision	recall	f1-score	support
hispanic	0.78	0.66	0.72	169,058
nh_black	0.66	0.25	0.36	149,299
nh_white	0.72	0.94	0.81	573,470
other	0.12	0.02	0.03	40,192
weighted_avg	0.69	0.72	0.68	959,848

Note: We decided on 2M records because of computational constraints and because the performance of the model trained on 2M records was very close to the performance of the model trained on 1M records.

SI 3 Performance of the Gradient Boosting Models

Table SI 3.1: Performance of the Last Name Gradient Boosted Trees model on the test set.

	precision	recall	f1-score	support
hispanic	0.78	0.67	0.72	38,961
nh_black	0.62	0.08	0.15	13,726
nh_white	0.56	0.92	0.7	60,970
other	0.35	0	0	14,374
weighted_avg	0.61	0.62	0.54	134,898

Table SI 3.2: Performance of the Full Name Gradient Boosted Trees model learned on 2M examples on the test set

	precision	recall	f1-score	support
asian	0.73	0.1	0.17	27,829
hispanic	0.75	0.53	0.62	169,058
nh_black	0.76	0.08	0.14	149,299
nh_white	0.67	0.96	0.79	573,470
other	0.43	0	0.01	40,192
weighted_avg	0.69	0.68	0.61	959,848

Note: We decided on 2M records because of computational constraints and because the performance of the model trained on 2M records was very close to the performance of the model trained on 1M records.

*

SI 4 Performance of the LSTM Models

Table SI 4.1: Performance of the Last Name LSTM model on the test set.

	precision	recall	f1-score	support
asian	0.41	0.19	0.26	6,859
hispanic	0.83	0.7	0.76	38,927
nh_black	0.59	0.4	0.48	13,719
nh_white	0.62	0.87	0.72	60,919
other	0.27	0.09	0.13	14,360
weighted avg	0.63	0.65	0.62	134,784

Table SI 4.2: Performance of the Full Name LSTM model on the test set.

	precision	recall	f1-score	support
asian	0.65	0.57	0.61	27,827
hispanic	0.79	0.83	0.81	169,038
nh_black	0.73	0.69	0.71	149,283
nh_white	0.85	0.9	0.88	573,409
other	0.39	0.08	0.13	40,187
weighted avg	0.8	0.81	0.8	959,744

SI 5 Performance of the Transformer Models

Table SI 5.1: Performance of the Last Name Transformer model on the test set.

	precision	recall	f1-score	support
hispanic	0.77	0.66	0.71	38936
nh_black	0.47	0.07	0.12	13717
nh_white	0.56	0.91	0.69	60910
other	0.00	0	0	14361
weighted	avg	0.55	0.61	0.53

Table SI 5.2: Performance of the Full Name Transformer model on the test set.

	precision	recall	f1-score	support
hispanic	0.66	0.51	0.57	169,041
nh_black	0.49	0.04	0.07	149,283
nh_white	0.66	0.94	0.78	573,408
other	0	0	0	40,188
weighted_avg	0.61	0.66	0.58	959,744