Can We Trust Race Prediction?

Cangyuan Li

July 10, 2023

Abstract

In the absence of sensitive race and ethnicity data, researchers, regulators, and firms alike turn to proxies. In this paper, I construct the most comprehensive database of first and surname distributions in the US in order to improve the coverage and accuracy of Bayesian Improved Surname Geocoding (BISG) and Bayesian Improved Firstname Surname Geocoding (BIFSG). Then, I present an ensemble model that, to the best of my knowledge, outperforms existing solutions, including BISG or BIFSG alone. The ensemble has greater than 90% accuracy for all classes and achieves higher precision, recall, and F1 scores than the literature, especially for minority classes. The ensemble consists of BISG, BIFSG, and a Bidirectional Long Short-Term Memory (LSTM) model trained on a novel dataset of voter registration data from all 50 US states. Finally, I seek to place a rough upper bound on the performance of models that rely only on name and location data by investigating the most ideal case–a simple lookup table.

Keywords: key1, key2, key3

JEL Codes: key1, key2, key3

1 Introduction

Race prediction has wide applications across a diverse range of fields, from lending to criminal justice to healthcare. As race and ethnicity are sensitive pieces of information, many datasets, public and private, do not have access to "true" race, and must therefore rely on proxies. For example, the Consumer Finance Protection Bureau (CFPB) uses BISG (name and zip code) in their fair lending analysis Bureau [2014]. Race prediction is essential to research involving racial outcomes, such as in [Brown et al., 2016, Frame et al., 2022, Clifford et al., 2023]. As such, an accurate proxy for race is paramount. However, data availability and generalizability are important considerations as well. Even in the absence of self-reported race, much better models exist. Image-based models such as Facenet512 achieve accuracies above 99%, even better than humans (98%). Furthermore, it is likely that additional or more granular features beyond the most generally available (name and zip code), such as income and address, would greatly improve accuracy. In this paper, I seek to place a rough upper bound on the performance of models that use only name and geography. First, I obtain a nationally representative corpus of names and zip code tabulation areas (ZCTAs) from L2 voter registration data. Additionally, I use public Paycheck Protection Program (PPP) data to create a clean dataset to validate against. To the best of my knowledge, no model incorporates this data into their training set, allowing me to conduct the fairest possible horse race. I then train a Bidirectional LSTM on the L2 data and show that it achieves higher precision, recall, and F1, than existing models. Then, I provided expanded first and last name tables for BISG and BIFSG, and show that an ensemble model outperforms any algorithm alone. Finally, I use the L2 data to create a lookup table and assess the performance of name and geography in the most ideal case.

1.1 Definitions

Throughout the paper, I use several metrics to assess the performance of different models. For each class, I calculate the number of True Positives (e.g. the predicted and self-reported races are both Asian), True Negatives (e.g. the predicted race is not Asian and the self-reported race is not Asian), False Positives (e.g. the predicted race is Asian but the self-reported race is not Asian), and False Negatives (e.g. the predicted race is not Asian but the self-reported race is Asian). Since each model returns the probability a person is of a certain race, I determine the predicted race by taking the maximum of the probabilities, which I call "Max". The following table provides a summary of the metrics used.

Metric	Formula	Interpretation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Ratio of correct predictions to total predictions
Precision	$\frac{TP}{TP+FP}$	Percent of correct positive predictions
Recall	$\frac{TP}{TP+FN}$	Percent of actual positives identified
F1 Score	$\frac{2\times Precision\times Recall}{Precision+Recall}$	Harmonic mean of precision and recall
Support		The number of samples that have a valid prediction
Coverage		The percentage of samples that have a valid prediction

2 Literature Review

This paper contributes to the race prediction literature, both in terms of model development and evaluation. Sood and Laohaprapanon [2018] and Xie [2022] both use Florida voter registration data to train an LSTM to predict race from first and last name. Voicu [2018] improves upon the BISG algorithm introduced by Elliot et al. [2009] by adding first name data. This paper's key contributions are to significantly improve the coverage and accuracy of BIFSG, provide a state-of-the-art machine learning model to fill in the gaps left by BIFSG, and make available a clean, nationally representative dataset for model developers to benchmark against.

3 Data

3.1 L2

I source 2023 voter data from L2, one of the leading providers of voter data in the US. The data spans 58 US states / territories, representing 32,034 out of 33,121 (96.7%) unique ZCTAs. I filter to the four major race / ethnicity categories: Non-Hispanic Asian, Non-Hispanic Black, Hispanic, and Non-Hispanic White. When self-reported race is not directly available, I use the ethnicity field, and follow the guidelines used by the Office of Management and Budget and the Census Bureau. For example, "White" includes people who report their ethnicity as German, Irish, English, etc, whereas "Asian" includes people who originate from countries such as China, India, and Japan Census [2022]. Table 1 provides a breakdown.

Table 1: L2 Racial Distribution

Race	Total	Self-Reported	From Ethnicity
Asian	7,722,316	575,210	7,147,106
Black	6,681,001	6,681,001	0
Hispanic	23,758,058	493,213	23,264,845
White	120,272,305	17,774,105	102,498,200

3.2 PPP

I use US Paycheck Protection Program (PPP) data to build a nationally representative database of name, geography (zip code), and self-reported race. I begin with a dataset of 11,460,475 loans spanning April 3, 2020 to May 31, 2021 and subset to the 1,211,770 name / zip pairs that both self-report race and are person names. Non-person names are removed using custom list of 1,000 filter words, such as "llc", "installation", and so on. The final dataset represents 1,066,605 unique first name / last name / ZCTA triplets, 27,702 out of 33,121 (84.6%) ZCTAs, and 57 states / territories.

Conducting a fair horse race is harder than it may first appear. The ultimate goal of these

models is to perform in a "real-world" setting. However, there are many equally valid "real-world" settings, and model performance can oscillate wildly between different distributions. For example, there is often significant overlap between Black and White names. A name such as "Dorothy Brown" encodes very little information about whether that person is Black or White, even to a human. Therefore, if Model X overpredicts Black, it would perform well on a sample where Black is the majority class, but would exhibit high false positive rates on a sample that is majority White. To address this issue, I assume that the target distribution roughly reflects the US population, and draw a nationally representative sample of 200,000 observations[1]. Table 2 provides a breakdown.

Table 2: PPP Racial Distribution

Race	Total	Sampled
Asian	57,303	12,202
Black	559,667	26,059
Hispanic	145,913	39,089
White	449,137	122,647

4 Models

4.1 BISG

The canonical version of BISG, developed by the Rand Corporation in 2009, calculates the probability a person is of a certain race as

$$P(r|s,g) = \frac{P(r|s) \times P(g|r)}{\sum_{r=1}^{6} P(r|s) \times P(g|r)}$$

where r is one of American Indian or Alaska Native, Asian or Pacific Islander, Black,

[1] I use the July 1, 2022, population estimates from `https://www.census.gov/quickfacts/fact/table/US/PST045222`–Asian: 5.9%, Black: 12.6%, Hispanic: 18.9%, White: 59.3%.

Hispanic, or Multiracial, s denotes surname, and g denotes geography. For the purposes of this paper, the geography is at the ZCTA level, although it can be defined at the census block, census tract, county, and state levels as well. $P(g|r)$ is the percentage of people of a certain race that live in the specified geography. $P(r|s)$ is defined as the percentage of people with a given surname that are of that race, and is calculated from the 2010 US census. The Census surname table comprises all surnames that appear more than 100 times, and yields 162,254 unique surnames covering 90% of the US population Comenetz [2016]. To improve coverage, I update the table with probabilities calculated from the L2 voter data, preferring the Census values if they exist. Since the L2 data is highly imbalanced, I draw a nationally representative sample, otherwise the probabilities would be artificially biased towards the majority classes. Additionally, I only consider the four major racial categories, allowing me to have substantially more observations. When building the table, I follow Tzioumis [2018] by deleting suffixes such as "JR", names that are only one character long, deleting blanks and hyphens, and only considering names that either have 30 or more observations or names that have 15-29 observations and represent one and only one race. I refer to this combined version as "iBISG." Tables 3 and 4 summarize their performances.

Table 3: BISG Stats (Max)

Race	Accuracy	Precision	Recall	F1 Score	Coverage	Support
Asian	0.986	0.904	0.842	0.872	0.875	10,674
Black	0.915	0.689	0.675	0.682	0.947	24,673
Hispanic	0.943	0.913	0.794	0.849	0.947	37,008
White	0.882	0.882	0.93	0.905	0.902	110,663

Table 4: iBISG Stats (Max)

Race	Accuracy	Precision	Recall	F1 Score	Coverage	Support
Asian	0.985	0.899	0.842	0.869	0.885	10,798
Black	0.915	0.69	0.676	0.683	0.951	24,777
Hispanic	0.943	0.91	0.794	0.848	0.951	37,176
White	0.882	0.882	0.929	0.905	0.908	111,314

iBISG shows modest improvements in coverage while showing virtually unchanged performance. That performance does not change is not surprising–the Census Bureau should have the highest quality data and the most observations to work with.

4.2 BIFSG

Voicu [2018] offers an extension of BISG in BIFSG, a similar algorithm that incorporates first name data. BIFSG calculates the probability a person is of a certain race as

$$P(r|s,f,g) = \frac{P(r|s) \times P(f|r) \times P(g|r)}{\sum_{r=1}^{6} P(r|s) \times P(f|r) \times P(g|r)}$$

All variables are defined in the same way as in BISG. The first name data comes from Tzioumis [2018], who uses Home Mortgage Disclosure Act (HMDA) data. I run the same routine as above to create the first name table from L2 voter data. Since the HMDA data has fewer observations, I prefer the L2 values if they exist. I refer to this combined version as "iBIFSG." Tables 5 and 6 summarize their performances.

Table 5: BIFSG Stats (Max)

Race	Accuracy	Precision	Recall	F1 Score	Coverage	Support
Asian	0.989	0.942	0.823	0.879	0.589	7,187
Black	0.941	0.7	0.616	0.656	0.536	13,957
Hispanic	0.949	0.928	0.797	0.858	0.764	29,881
White	0.904	0.902	0.961	0.931	0.843	103,437

Table 6: iBIFSG Stats (Max)

Race	Accuracy	Precision	Recall	F1 Score	Coverage	Support
Asian	0.988	0.905	0.873	0.889	0.764	9,325
Black	0.936	0.713	0.766	0.738	0.786	20,487
Hispanic	0.947	0.911	0.813	0.859	0.881	34,433
White	0.908	0.917	0.939	0.928	0.887	108,832

In this case, iBIFSG shows significant improvements in coverage across all classes, with the greatest gains coming from minority classes. In particular, iBIFSG achieves a 12.5% increase in F1 Score and a 46.6% increase in coverage compared to BIFSG for Black, traditionally one of the hardest to predict classes. The gains are mainly due to increased recall. That is to say iBIFSG correctly identifies 76.6% of all Black borrowers in the sample, compared to just 61.6% for BIFSG.

4.3 LSTM

While iBIFSG shows good performance, and already represents a significant step forward in terms of coverage, it still struggles with missing data. Furthermore, names that do not appear in the probability files algorithms such as BISG and BIFSG rely on often are correlated with nationality. For example, many African and Eastern European names, such as "Jurczewsky", "Semuyaba", and "Ng'ethe", to name a few, do not appear in the files. Users relying on such algorithms may be systemically excluding certain groups from their sample. Therefore, I train a Bidirectional LSTM to address these gaps. Bidirectional LSTMs add a backward layer to a regular LSTM where the information is reversed, and have been shown to be able to better capture the context of text Graves and Schmidhuber [2005]. The base model was trained using Keras and consists of an embedding layer with an embedding dimension of 256, four LSTM layers with hidden size 512 and a dropout rate of 0.2, and a final dense layer with softmax activation. I use the Adam optimizer with .001 learning rate, and character-encode names before passing them to the embedding layer. Character-level features work well with neural networks since they are good at extracting information from raw data Zhang et al. [2015].

Before training, I undersample the dataset so that each class has an equal number of observations. In comparison, an imbalanced dataset could lead the model to optimize by simply predicting the majority class (White) most of the time. To the best of my knowledge, the resulting model has better performance than existing models in the literature. For

example, my model achieves a F1 score of 0.639 for Black, compared to 0.552 for Sood and Laohaprapanon [2018], 0.513 for Xie [2022], and 0.47 for Kotova [2021]. A full comparison against Sood and Laohaprapanon [2018] (ethnicolr) and Xie [2022] (rethnicity) on the aforementioned PPP test sample can be found in tables 12 and 11[2]. Importantly, the errors the model makes seem to be "reasonable"–one could imagine a human making the same mistake with the same information. For example, the model gets "Felicia Gray", "Barbara Middleton", and "Karen Ross" wrong, who all self-report as Asian. Similarly, the model predicts "Surinder Kaur", "Balbir Ghandi", and "Tu Vuong" as Asian, but they self-report as White.

Table 7: First-Last Stats (Max)

Race	Accuracy	Precision	Recall	F1 Score	Coverage	Support
Asian	0.975	0.772	0.839	0.804	1.0	12,202
Black	0.898	0.589	0.705	0.642	1.0	26,059
Hispanic	0.937	0.871	0.795	0.831	1.0	39,089
White	0.862	0.896	0.876	0.886	1.0	122,647

Location also encodes important information. While more granular location data would be ideal (such as census tract or even address), the most common models typically use zip code, as it is the most readily available. Instead of adding location features to the model, which may not be portable (if, for example, the racial distribution of a ZCTA changes drastically in the future, or a user wants to use tract-level data), I use the following equation:

$$P(r|n,g) = \frac{P(r|n) \times P(g|r)}{\sum_{r=1}^{4} P(r|n) \times P(g|r)}$$

,

where n is name, and $P(r|n)$ are the probabilities returned by the aforementioned name-only model. The other terms are exactly as in BISG. The resulting model achieves similar

[2]Kotova [2021] does not have an associated open-source package to test against.

results for Asian and Hispanic, but makes a significant leap for Black (a 17% increase in F1 score) and a modest gain for White (a 3% increase in F1 score). This makes sense, as Black and White names are the most likely to be confused for each other, and is where location features can make the most difference. For instance, a person with the surname "Li" is likely Asian regardless of location, and indeed location may even just add noise in such cases.

Table 8: First-Last-ZCTA Stats (Max)

Race	Accuracy	Precision	Recall	F1 Score	Coverage	Support
Asian	0.974	0.744	0.864	0.799	0.994	12,130
Black	0.925	0.664	0.854	0.747	1.0	26,046
Hispanic	0.941	0.881	0.805	0.841	0.999	39,052
White	0.893	0.934	0.888	0.911	1.0	122,601

4.4 Ensemble

In this section, I present a simple ensemble model that maximizes coverage while maintaining high performance. I take the weighted average of the predictions made by iBIFSG, iBISG, and First-Last-ZCTA, assigning equal weight to each model that is able to make a prediction. Table 9 reports the performance.

Table 9: Ensemble Stats (Max)

Race	Accuracy	Precision	Recall	F1 Score	Coverage	Support
Asian	0.979	0.803	0.862	0.831	1.0	12,202
Black	0.931	0.706	0.811	0.755	1.0	26,059
Hispanic	0.943	0.895	0.802	0.846	1.0	39,089
White	0.9	0.921	0.915	0.918	1.0	122,647

The ensemble reports higher F1 scores across the board than First-Last-ZCTA and maintains perfect coverage. In the future, more sophisticated weighting schemes may improve performance. For example, it is possible that certain models perform well in certain geographies, and geography-specific weights could be constructed. However, since the PPP

sample does not cover all ZCTAs, and does not have a statistically significant number of observations for every ZCTA, I do not attempt this exercises.

4.5 Lookup Table

Finally, I investigate the ideal case by creating a lookup table from the L2 data. If one had perfect information, i.e. knew all the names and races of everyone living in every ZCTA, the optimal prediction for a given name and ZCTA is simply the count of people in that ZCTA with that name that are Asian, Black, Hispanic, or White divided by the total number of people with that name. However, since the L2 data is not sufficiently large (especially for Asian and Black) to calculate probabilities at the ZCTA-level, I instead create the lookup table based on first and last name alone and incorporate ZCTA information using naive Bayes. The same name cleaning and filtering procedures as described in 4.1 are applied to the full name (first name + last name). Then,

$$P(r|n,g) = \frac{P(r|n) \times P(g|r)}{\sum_{r=1}^{4} P(r|n) * P(g|r)}$$

Table 10 reports the performance of the lookup table.

Table 10: Lookup Stats (Max)

Race	Accuracy	Precision	Recall	F1 Score	Coverage	Support
Asian	0.993	0.974	0.917	0.944	0.331	4,034
Black	0.933	0.683	0.629	0.655	0.231	6,029
Hispanic	0.955	0.936	0.891	0.913	0.4	15,622
White	0.904	0.898	0.937	0.917	0.274	33,593

Compared to iBIFSG, the lookup table has much poorer coverage (as expected), but exhibits higher F1 scores for Asian (6.2%), and Hispanic (6.2%). However, Black (-11%) and White (-1.2%) actually regress. A few explanations are possible. One is that the statistics

for the lookup table are more noisy since it is only able to make predictions for a small number of observations. Another is that the names that have enough samples (at least 30) to appear in the data are names that are neither uniquely Black nor White. That is to say, there are more ambiguous names like "Dorothy Brown" than relatively more clear-cut names like "Onyiuke Anthonia" due to lack of data.

5 Discussion

Can we trust race prediction? Even with perfect information, using only name and location inherently limits the accuracy of available models. For instance, a Black person with an ambiguous name ("Barbara Jackson", "Ashley Jackson") living in New York City will likely always be mislabeled by such models as White. Similarly, many Filipinos with Hispanic-sounding names (such as "Maria Cruz Santos") will be labeled as Hispanic. However, it is clear that models have the ability to achieve very high performance. iBIFSG and the resulting ensemble already show significant improvements over existing solutions that have been used in such critical contexts as fair lending analysis, and more data or advances in architectures could allow models to approach an acceptable upper bound in performance. For instance, Hu et al. [2021] use a Dual-LSTM architecture to improve gender classification. Voter data is inherently not representative, as not everybody is registered to vote, especially Blacks and Hispanics Sood and Laohaprapanon [2018]. A waterfall architecture, where separate models are trained with respect varying levels of granularity (for example, first name + last name + income + tract $\rightarrow$ first name + last name + ZCTA $\rightarrow$ first name + last name), and combined at the end, could maximize both performance and coverage.

References

D. P. Brown, C. Knapp, K. Baker, and M. Kaufmann. Using bayesian imputation to assess racial and ethnic disparities in pediatric performance measures. *Health Services Research*, 2016. doi: https://doi.org/10.1111/1475-6773.12405.

C. F. P. Bureau. Using publicly available information to proxy for unidentified race and ethnicity. Technical report, Consumer Finance Protection Bureau, 2014.

Census. Census bureau race definitions. https://www.census.gov/topics/population/race/about.html, 2022.

C. P. Clifford, W. C. Gerken, and T. Qiu. Racial concordance in the market for financial advice. *The Review of Corporate Finance Studies*, 2023.

J. Comenetz. https://www2.census.gov/topics/genealogy/2010surnames/surnames.pdf, 2016.

M. N. Elliot, P. A. Morrison, A. Fremont, D. F. McCaffrey, P. Pantoja, and N. Lurie. Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 2009. doi: https://doi.org/10.1007/s10742-009-0047-1.

W. S. Frame, R. Huang, E. J. Mayer, and A. Sunderam. The impact of minority representation at mortgage lenders. Working Paper 30125, National Bureau of Economic Research, June 2022. URL http://www.nber.org/papers/w30125.

A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2005.06.042. URL https://www.sciencedirect.com/science/article/pii/S0893608005001206. IJCNN 2005.

Y. Hu, C. Hu, T. Tran, T. Kasturi, E. Joseph, and M. Gillingham. What's in a name? - gender classification of names with character based machine learning models. *CoRR*, abs/2102.03692, 2021. URL https://arxiv.org/abs/2102.03692.

N. Kotova. A deep learning approach to predicting race using personal name and location (natural language processing). 2021.

G. Sood and S. Laohaprapanon. Predicting race and ethnicity from the sequence of characters in a name, 2018.

K. Tzioumis. Demographic aspects of first names. *Scientific Data*, (180025), 2018.

I. Voicu. Using first name information to improve race and ethnicity classification. *Statistics and Public Policy*, 5(1):1–13, 2018. doi: 10.1080/2330443X.2018.1427012. URL https://doi.org/10.1080/2330443X.2018.1427012.

F. Xie. rethnicity: An r package for predicting ethnicity from names. *SoftwareX*, 17:100965, 2022. ISSN 2352-7110. doi: https://doi.org/10.1016/j.softx.2021.100965. URL https://www.sciencedirect.com/science/article/pii/S2352711021001874.

X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

Tables

Table 11: Rethnicity Stats (Max)

Race	Accuracy	Precision	Recall	F1 Score	Coverage	Support
Asian	0.939	0.502	0.882	0.64	1.0	12,202
Black	0.811	0.386	0.763	0.513	1.0	26,059
Hispanic	0.926	0.845	0.761	0.801	1.0	39,089
White	0.763	0.909	0.682	0.779	1.0	122,647

Table 12: Ethnicolr Stats (Max)

Race	Accuracy	Precision	Recall	F1 Score	Coverage	Support
Asian	0.974	0.897	0.643	0.749	1.0	12,202
Black	0.901	0.673	0.469	0.552	1.0	26,059
Hispanic	0.932	0.882	0.754	0.813	1.0	39,089
White	0.845	0.828	0.943	0.882	1.0	122,647