Classifying Gender Using SSA Name Data

Katherine Hovland
Marquette University
Milwaukee, WI 53233
katherine.hovland@marquette.edu

Abstract

Because first names are a common identifier, the ability to draw conclusions about an individual based on first name is a valuable capability. The SSA database provides the opportunity to draw conclusions about the gender of individuals who are members of the living United States population. This paper demonstrates that Naïve-Bayes classification can be used to classify gender with high levels of accuracy and decade of birth with low accuracy.

1 Introduction

The goal of this project was to create a classification model that can predict gender and approximate year of birth based on first name, using data from the Social Security Administration (SSA). Such a model has many practical applications, allowing individuals or organizations to draw conclusions about individuals given only their first name. For example, a company could utilize gender classification to analyze a customer database and provide targeted advertising.

Figure 1 shows four examples of how the popularity of a given name varies based on age and year of birth. The name "Deborah," for example, is much more popular for women than for men and spiked in popularity in the 1950s. A person named Deborah is thus likely to be female and in their sixties.

This project looks exclusively at the United States and focuses specifically on the current living population. This distinction is important both for predicting birth year and for more accurately predicting gender. Figure 1 shows how the gender distribution of certain names changes over time. For example, "Jackie" was once a predominantly male name but became more popular for women during the 1950s.

Figure 1 also reveals some of the potential problems with using first name to predict gender and age. Some names, such as "Riley" and "Jackie," do not show a clear gender split, and thus are more likely to be misidentified. Likewise, while some names such as "Deborah" show a clear spike in a single decade, others, such as "John," are popular in multiple decades and remain common across the entire range of years. These problems limit the effectiveness of the model and will be discussed further below.

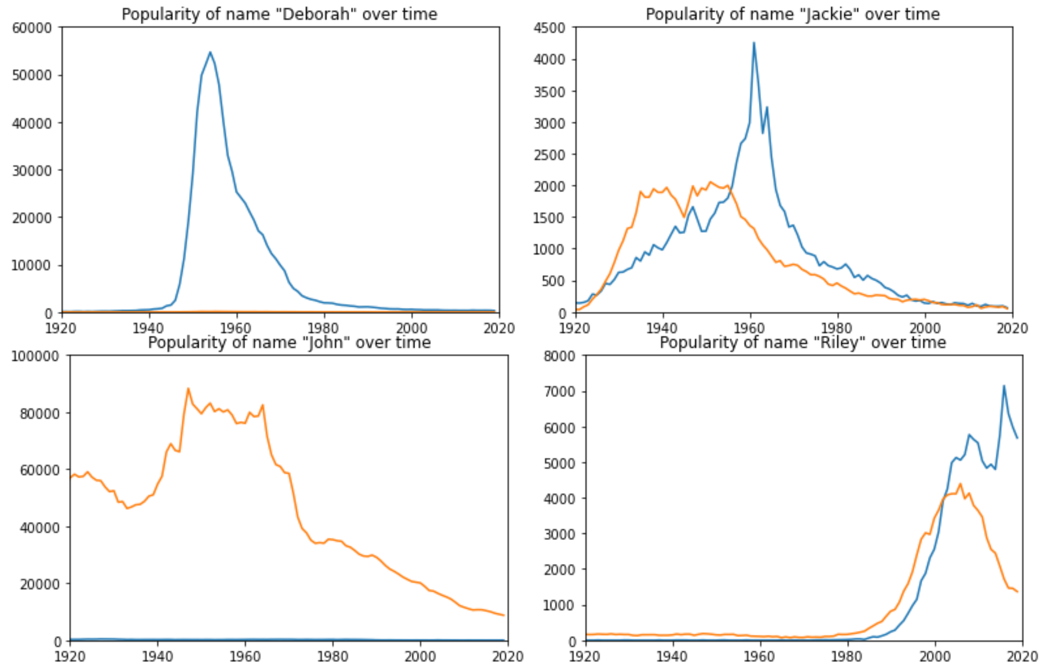

Figure 1: Change in popularity of given names "Deborah," "Jackie," "John," and "Riley" over time. Popularity for male babies is depicted in orange and popularity for female babies is in blue. Note that the scale of the y-axis is different for each graph, and this figure should not be used to compare one name to another.

2 Lit Review

The authors of a paper discussing the gender of authors of scientific papers (West et.al., 2013) used the top 1000 names in the SSA dataset to categorize names as male, female, or unisex. This methodology valued high degrees of accuracy for a limited number of names, and thus did not attempt to utilize the full dataset or classify more challenging names. Because it analyzed historical data, it did not adjust the dataset to reflect the living population. "Jane, John … Leslie? A Historical Method for Algorithmic Gender Prediction" (Blevins & Mullen, 2015) incorporates additional variables, including approximate year of birth, to more accurately identify names in historical data.

Research by Gallagher and Chen (2008) used the top 1000 names to improve the identification of name and age in image classification. Their model also used name and gender information to assign names to individuals in a photograph. Their work demonstrates the value of using names to predict age and gender.

Other methods of gender classification (Larivière et al., 2013; Strømgren, 2015; Wais, 2016) exist which offer highly accurate gender classification in a wide variety of circumstances. However, they often involve multiple data-sources or variables and may not be appropriate in all contexts.

3 Dataset

3.1 Social Security Administration Data

For every year since 1880, the SSA provides a list of the number of babies born in the United States with a given name, broken down by gender. The dataset recognizes two genders, male and female, based on the sex assigned to an individual at birth.

In order to protect privacy, the list excludes names with fewer than five occurrences in that year. The dataset also only covers babies who were born in the United States and applied for a social security number. It does not reflect emigration or immigration. Regardless, the SSA dataset is quite comprehensive and reflects a large portion of the population of the United States. This project used data from the years 1900-2019.

3.2 Living Population

The percentage of people born in a given year who are still alive today was used to approximate the number of people alive with a given name. The percentage living was calculated by *24/7 Wall St.* (Comen, 2020) using US Census Bureau estimates of the native population by age and Centers for Disease Control and Prevention information on birth year. The data covers the birth years 1936-2019.

Because 24/7 Wall St. did not provide data on the years 1900-1935, the data that was given was used to extrapolate the past data using linear regression. Since the relationship between birth year and living population is not linear, a third order polynomial basis function was applied to the data to improve fit. Linear regression on an 80-20 train-test split resulted in an R-squared score of 0.98014 for the train data and an R-squared score of 0.95739 for test data.

Extrapolation was then used to predict the percentages of the population born in the years 1900-1935 who are still alive. Figure 2 shows the estimated percentage of people born in a given year who are alive today. Extrapolation is typically inaccurate, and the corner point visible in Figure 2 is not a realistic expectation for this dataset. However, extrapolation provides a better estimator than 0. Caution should nonetheless be used if attempting to classify a population with a larger number of individuals who are 80 years or older.

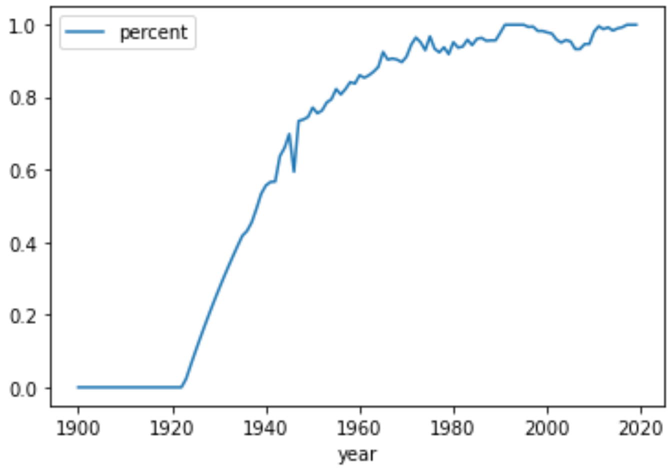

Figure 2: Percentage of the population alive today

This data was applied to the SSA data by multiplying the number of occurrences of each name in a given year by the percentage of the population born in that year alive today. It is important to note that this population estimate was applied equally to all names and did not consider factors such as gender or ethnicity that may impact life span.

4 Analysis

The analysis of this data utilized sklearn.naive_bayes.CategoricalNB to perform a Naïve-Bayes classification on representative samples of the SSA data. The use of representative samples, although less accurate than the full dataset, made the process of applying Naïve-

Bayes classification simpler.

4.1 Classifying by Gender

5 million male and 5 million female datapoints were chosen using the known frequencies of each name as weights. This representative sample was divided into train and test data using a 75-25 split.

The input data for the Naïve-Bayes classification was the names, with each name representing a single category, signified by a unique integer. The output variable to be predicted was gender, represented as "M" or "F."

A Laplacian correction was necessary to prevent a zero-frequency problem. One male and one female instance of each name was added to the training set. Real-life applications should anticipate the fact that not all names are included in the SSA dataset.

4.2 Classifying by decade

Similar methodology was used to classify the decade in which a person was likely born. Each decade was defined as the year 1900-1909, 1910-1919, and so on up to 2010-2019. The number of datapoints chosen for each decade was found by multiplying 1 million by the percentage of a people born in the first year of the decade, resulting in a representative sample of approximately 7 million datapoints. Although it resulted in uneven classes, this methodology was chosen to better represent the living population.

As above, the input data for the Naïve-Bayes classification was the names, with each name representing a single category, signified by a unique integer. The output variable to be predicted was the decade of birth, represented by the first year of the decade ("1900," "1910," etc.).

To avoid diluting the dataset, one instance of each name was added to the training set with the output variable "0." This additional class reduced the effectiveness of the model but provided an easy and efficient way to prevent the zero-frequency problem.

5 Results

5.1 Classifying by gender

The Naïve-Bayes classification by gender yielded an R-squared score of 0.96660 for the train data and an R-squared score of 0.97764 for the test data.

When using the model to predict gender, 55,897 out of a total of 2,500,000 test points were mislabeled, or approximately 2.235%. Table 1 provides the confusion matrix for a more detailed breakdown of the predicted vs. actual classification.

Table 1: Gender Classification Confusion Matrix

		Actual	
		Female	Male
Predicted	Female	1,219,643	25,204
	Male	30,693	1,224,460

The high accuracy scores and reasonably well-balanced confusion matrix suggest that this model can be used to reliably predict gender in most circumstances. Without additional data about an individual, some misclassifications are inevitable, but the model will accurately identify the gender of about 97% of individuals in a representative sample.

5.2 Classifying by decade

The Naïve-Bayes classification by decade yielded an R-squared score of 0.33964 for the train data and an R-squared score of 0.3276 for the test data. These scores, though low, are higher than random chance would yield.

When using the model to predict decade, 1,219,352 out of a total of 1,813,664 test points were mislabeled, or approximately 67%. However, Figure 3 reveals that many of those misclassifications may be off by only one or two decades by comparing the distribution of actual decades for select predicted decades.

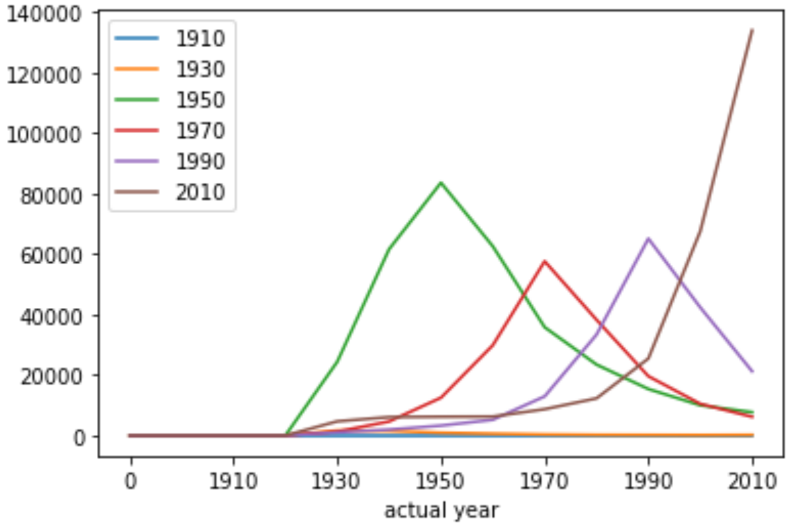

Figure 3: Number of names from each actual decade. Each color represents all individuals predicted to have been born in a given decade.

The low accuracy scores indicate that this model cannot accurately identify the decade of birth of an individual by first name. However, it may still provide some insight into the approximate age of individuals. Larger intervals, perhaps for predicting generation, would likely be more accurate.

6 Conclusion

This project successfully produced a model for identifying gender in living populations in the United States. Although it performed satisfactorily, it has not been proven effective with a sample that is not representative of the living United States native population.

It was less successful at producing a model to identify the decade in which an individual was born. This model could likely be improved by incorporating other variables, using a different technique to prevent the zero-frequency problem, or using a larger interval for the classes.

Overall, both models provide a simple technique for classifying individuals using only their first names, with varying degrees of accuracy.

References

Blevins, C, & Mullen, L. (2015).Jane, John ... Leslie? A historical method for algorithmic gender prediction. *Digital Humanities Quarterly*, 9(3), 17-35. http://www.digitalhumanities.org/dhq/vol/9/3/000223/000223.html

Comen, E. (2020, July 29). *How many people are alive from the year you were born?* 24/7 Wall St. https://247wallst.com/special-report/2020/07/29/how-many-people-are-alive-from-the-year-you-were-born/10/

Gallagher, A. C. & Chen, T. (2008). Estimating age, gender, and identity using first name priors. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1-8. https://ieeexplore.ieee.org/abstract/document/4587609

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: global gender disparities in science. *Nature*, 504(7479), 211–213. https://www.nature.com/news/bibliometrics-global-gender-disparities-in-science-1.14321

Social Security Administration. *Beyond the top 1000 names*. https://www.ssa.gov/oact/babynames/limits.html

Strømgren, C. (2015). genderize.io. URL http://genderize.io.

Wais, K. (2016). Gender prediction methods based on first names with genderizeR. *The R Journal*, 8(1), 17-37. https://journal.r-project.org/archive/2016-1/wais.pdf

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The role of gender in scholarly authorship. *PLOS ONE*, 8(7), 17-20.