Enhancing Minority Business Representation through Machine Learning: A Case Study in Fairfax County

CS4991 Capstone Report, 2025

Anjali Mehta
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
wfn7ad@virginia.edu

ABSTRACT

Minority-owned businesses are often underrepresented in existing datasets, creating challenges for policymakers aiming to foster economic diversity. We addressed this issue by developing a binary classification model to better identify minority-owned businesses within Fairfax County, Virginia. Using data from various open-source datasets and external verifications, we applied machine learning techniques, such as natural language processing (NLP) tools like RaceBERT, to analyze executive names, company names, and location demographics. The model successfully increased the estimated representation of minority-owned businesses from 7% to 41.75%, significantly reducing errors in existing datasets. Challenges included data limitations, potential biases in demographic inferences, and the small size of the training set. Future work will focus on expanding the dataset, improving classification accuracy, and applying the model to other regions to support inclusive policymaking and promote business diversity.

1. INTRODUCTION

Tracking economic diversity is essential for policymakers and outreach programs aiming to support minority-owned businesses. A minority owned business, as defined by U.S. standards, is one in which at least 51% of the ownership belongs to individuals from socially and economically disadvantaged minority groups (2024). However, existing datasets often fail to accurately identify these businesses, limiting efforts to promote equitable economic development. Microdata, which allows for the study of business activity at small geographic levels, presents an opportunity to improve these tracking efforts.

A case study conducted in Fairfax County, Virginia, highlights this issue. The Annual Business Survey (ABS) estimated that approximately 38% of businesses were minority owned in 2017, whereas Mergent Intellect, our primary dataset, reported only 7%. While Mergent Intellect's methodology for data collection remains unclear, our findings suggest that it primarily includes only officially registered minority-owned businesses, which would omit a significant portion of unregistered ones. This underrepresentation raises an important question: How are minority-owned

businesses distributed across Fairfax County geographically?

2. RELATED WORKS

Existing research highlights key challenges in identifying and analyzing minority-owned businesses. Minority entrepreneurs are underrepresented in business datasets, leading to significant difficulties in obtaining accurate microdata (Puryear et al., 2019). Traditional data collection methods, such as self-reported surveys and business registries, often fail to capture unregistered or informally operated minority-owned businesses, resulting in incomplete datasets. This highlights the need for refined data collection approaches that incorporate external verification methods and more inclusive strategies to improve representation.

Advancements in machine learning (ML) and natural language processing (NLP) offer promising solutions for improving demographic identification in business datasets. RaceBERT, an NLP model designed to infer race and ethnicity based on names, has been shown to enhance demographic classification accuracy (Parasurama, 2021). However, the use of ML models for demographic inference raises concerns about potential biases, ethical implications, and privacy risks, particularly when applied to sensitive demographic data. Hanna et al. (2025) highlight the risks associated with algorithmic bias in Artificial Intelligence (AI)/ML models, emphasizing the importance of designing classification models that minimize bias.

3. PROJECT DESIGN

This section outlines the methodology used to identify minority-owned businesses in Fairfax County, which involved aggregating and validating data from multiple sources, addressing misclassification issues in existing datasets such as Mergent Intellect, and developing a predictive model based on executive names, company language, and location demographics to enhance classification accuracy

3.1 Methodology

My methodology focuses on identifying minority-owned business in Fairfax County using data aggregation, validation, and predictive modeling. The overall process is shown in the Figure 1 below:

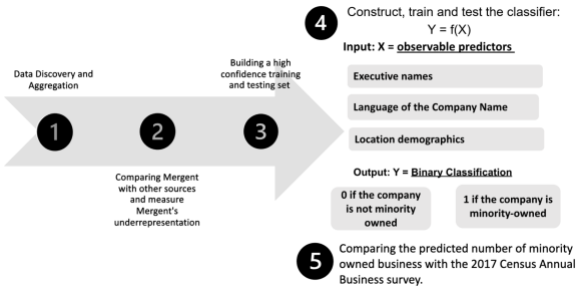

Figure 1: Methodology

3.1.1 Data Discovery and Aggregation

I compiled data from several minority-focused websites, as shown in Figure 2 below. The primary dataset was Mergent Intellect, but this dataset underrepresented minority-owned businesses in Fairfax County. I cross-referenced Mergent Intellect with the curated listing of high confidence minority-owned businesses.

Listings	Data Collection Method	Sample Size
Yelp	Consumer/Owner report businesses	871
Small Businesses and Supplier Diversity (SBSD)	Administrative Record	987
Data Axle	Census and public records	650
Chamber of Commerce	Businesses register + pay membership fee	435

Figure 2: Data Sources in Listing

Also, the Mergent Intellect data points that included executive's names reported minority ownership were also added to the listing. Figure 3 below demonstrates that although the Mergent Intellect dataset seemed large, I was only able to pull 743 usable data points from it.

Mergent Intellect	Number of Businesses
no filter	166K
w/executive names reported	12K
w/executive names reported + minority ownership	743

Figure 3: Usable Data Points Extracted from Mergent Intellect

Fuzzy matching was used to compare Mergent Intellect data with the Listing, resulting in only 70 overlapping businesses. This highlights the need for further data collection to improve accuracy.

3.1.2 Mergent's Misclassification

The issue with the Mergent Intellect data is that it underrepresents the number of minority businesses. I investigated the businesses that were classified as minority-owned with alternative sources and detected that 39.47% had been misclassified, as shown in Figure 4 below:

		Included in the additional data sources?		Total
		Yes	No	
Mergent Intellect data	Minority-owned	23	743	771
	Non-minority owned	15	17,413	17,428
Total		38	18,156	18,199

Figure 4: Mergent Intellect's Misclassification

3.1.3 Building Our Training and Testing Set

I created a balanced training set by cross-referencing Mergent Intellect data with the Listing. To reduce bias, I used "Listed" and "Not Listed" as proxies for minority and non-minority ownership. This bias could exist because, although I have high confidence that Listed businesses are minority-owned, I do not have confidence that Not-Listed businesses are not minority owned.

The final sample comprises 138 businesses. After finding the unique businesses labelled as "Not Listed," I extracted the same number of businesses classified as "Listed" by randomly sampling from the subset of 813 companies. This process is shown in Figure 5 below:

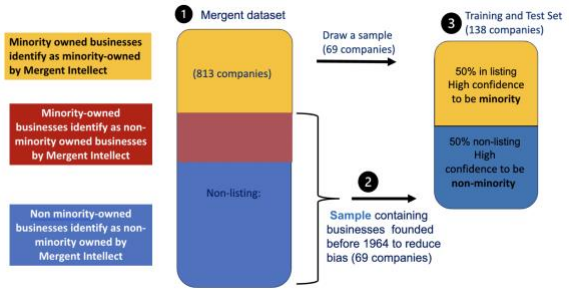

Figure 5: Process of Creating the Training and Test Set

3.2 Constructing the Classifier

The model uses three predictors: Business executive's names, language of the company name and proportion of non-minority individuals at the business location.

3.2.1 Executive's Names

Utilizing NLP tools such as RaceBERT and Ethnicolr, I created a prediction column that flags each name possessing a higher probability of being non-white (greater than 0.5) to belong to a member of a minority group. A value of 1 means an owner is predicted to be a minority and 0 if otherwise.

To visualize this distribution, I created the box plots shown in Figure 6 below. The box plot presents the probability scores of names classified as minority (1) and those classified as non-minorities (0).

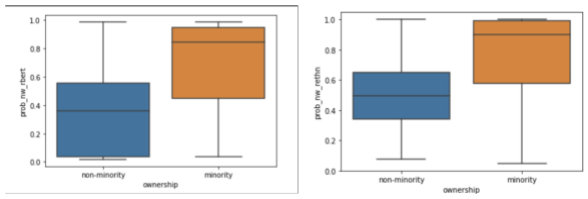

Figure 6: Box Plots of Executive's Names

3.2.2 Company Name

I analyzed company names using LangDetect and SpaceY to identify minority-spoken languages, such as Spanish or Arabic. Names in these

languages were flagged as minority (1). The distribution based on this indicator is shown in Figure 7 below:

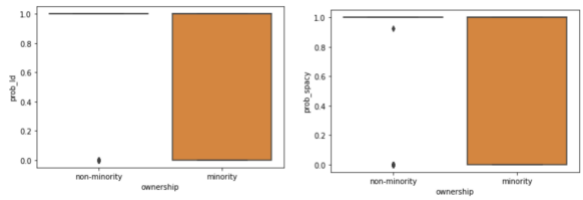

Figure 7: Distribution Based on Company Name

3.2.3 Company Location

Using 2021 census data, I analyzed the demographic composition of Fairfax County census tracts. Businesses in tracts with higher non-white populations were weighted accordingly.

3.3 Model Evaluation

The training set was split up and 30% of it was subsetted into the testing set. The model was run 10,000 times to produce evaluation metrics of the model, based on a confusion matrix.

3.4 Model Application

I predicted the minority status of businesses reported in the overlap of Mergent Intellect and the listing. Therefore, I estimate the proportion of businesses flagged as non-minority by Mergent Intellect, for which both the model and outside sources flagged as minority-owned. This process allows me to estimate how the model has reduced the error made by Mergent Intellect in flagging minority-owned businesses and non-minority owned businesses.

4. RESULTS

I developed a decision tree model to classify businesses, focusing on identifying minority-owned ones, as shown in Figure 8 below:

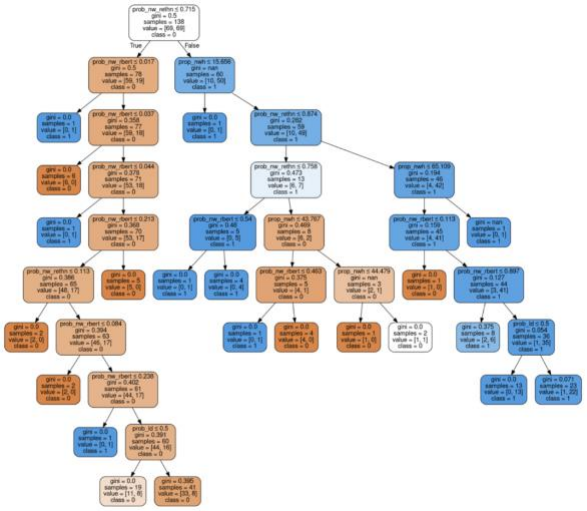

Figure 8: Decision Tree

Initially, Mergent's data had a 39% error rate, with only 7% of minority-owned businesses correctly classified. Applying the model to non-flagged businesses in the Mergent dataset reduced the error rate to 22.34% and increased the percentage of identified minority-owned businesses to 41.75%. While benchmarking against the 2017 Census (38% minority ownership in Fairfax County), the higher percentage is plausible given recent data.

	Percentage minority business in dataset	Percent Misclassified
Mergent Intellect Database	7%	39.47 %
After using Decision Tree Model	41.75 %	22.34 %

Figure 9: Results of Decision Tree

Due to a small sample size (138 businesses), the model was trained and tested 10,000 times, splitting data into 70% training and 30% testing sets. Performance metrics, averaged for evaluation, are shown in Figure 10 below:

Model Evaluation Metrics		
Average accuracy from the classifier	Average precision from the Classifier	Average F1-score from the Classifier
74.10%	70.11%	69.67%

Figure 10: Decision Tree Model Evaluation

5. CONCLUSION

Our model predicts that 41.75% of businesses in Fairfax County are minority-owned, a significant increase from the 7% reported by Mergent Intellect. This improvement highlights the potential of machine learning and natural language processing (NLP) tools, such as RaceBERT, to address underrepresentation in existing datasets. By leveraging executive names, company language, and location demographics, we were able to refine classification accuracy and provide a more inclusive picture of minority business ownership. These findings are critical for policymakers and community programs aiming to foster economic diversity and support underserved businesses. Accurate data enables targeted initiatives that promote equitable growth, strengthen local economies, and prioritize representation for historically marginalized groups.

The small size of our training set and the challenges of data collection, such as incomplete business listings and reliance on census-based demographic inferences, underscore the need for further validation and refinement. Despite these constraints, our model demonstrates the value of combining multiple data sources and advanced analytical techniques to improve the accuracy of minority business identification. Moving forward, expanding and validating our dataset will be essential to ensure the model's reliability and applicability to broader contexts.

6. FUTURE WORK

The next phase of this project will focus on expanding and validating our dataset to improve the model's accuracy and generalizability. Currently, the training set is limited to 138 businesses, which restricts the model's ability to capture the full diversity of minority-owned businesses in Fairfax County. To address this, we will conduct more extensive web scraping and data aggregation to identify additional businesses, particularly those that are unregistered or informally operated. This will help us create a more robust and representative dataset, reducing potential biases and improving the model's performance.

Additionally, we plan to explore the geographic and industry-specific distribution of minority-owned businesses within Fairfax County. Understanding where and in which sectors these businesses are concentrated can inform targeted policy interventions and resource allocation. Finally, we aim to apply our classifier model to other regions covered by the Social Impact Data Commons, such as the broader National Capital Region. This expansion will test the model's scalability and provide valuable insights into minority business ownership across different geographies, ultimately supporting more inclusive and data-driven policymaking.

REFERENCES

NMSDC. (2024, October 8). *Definition of an MBE*. https://nmsdc.org/certifications/definition-of-an-mbe/

Hanna, M., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., & Rashidi, H. (2025). Ethical and bias considerations in Artificial Intelligence/Machine Learning. *Modern Pathology*, *38*(3), 100686. https://doi.org/10.1016/j.modpat.2024.100686

Parasurama, P. (2021, December 9). *Racebert—A transformer-based model for predicting race and ethnicity from*

names. arXiv.org. https://arxiv.org/abs/2112.03807

Puryear, A., Rogoff, E., Lee, M.-S., Heck, R., Grossman, E., Haynes, G. & Onochie, J. (2019). Sampling minority business owners and their families: The understudied entrepreneurial experience. *Journal of Small Business Management*, *46*(3), 422–455. https://doi.org/10.1111/j.1540-627x.2008.00251.x