Business Diversity

Measuring Minority Business Ownership in Fairfax County, VA for the Social Impact Data Commons

Data Science for the Public Good
Social and Decision Analytics Division
Biocomplexity Institute

30 June 2023

University of Virginia

Project Team

DSPG, University of Virginia, Biocomplexity Institute, Social and Decision Analytics Division

- Anjali Mehta, Third Year @ UVA (CS Engineering)
- Prashanth Wagle, Grad Student @ UVA (CS)
- Trinity Chamblin, Second Year @ UVA (Stats & CS)

University of Virginia, Biocomplexity Institute, Social and Decision Analytics Division

- Aaron Schroeder, Research Associate Professor
- Joel Thurston, Senior Scientist
- Guy Leonel SIWE, Postdoc Researcher
- Treena Goswami, Postdoc Researcher

Fairfax Stakeholders

- Michelle Gregory, Fairfax Countywide Data Analytics Coordinator
- Stephen Tarditi, Director, Market Intelligence at Fairfax County Economic
- Scott Sizer, Fairfax Department of Economic Initiatives

This project has been sponsored by Mastercard Center for Inclusive Growth

Project Description

- A **Data Commons** is an open knowledge repository co-locating publicly available datasets alongside administrative records, advanced analytics, and visualization tools.
- Our aim is to understand the landscape of **economic diversity** -- focusing on minority-owned businesses within Fairfax County, Virginia

How the Census Defines a Minority Owned Business:

A U.S. company that is at least 51% owned by, and whose management and daily company operations are controlled by, one or more members of a socially and economically disadvantaged minority group (based on race)

Motivation

Our stakeholders are interested in answering this question:

How are minority businesses distributed across Fairfax County geographically?

How will we help?

- We will develop a predictive model to reduce the underrepresentation of minority business ownership
- We will compare the results of our model with that of the latest Census Annual Business Survey

Our Primary Dataset

There are many sources of data that enable the study of business activities at small geography levels...

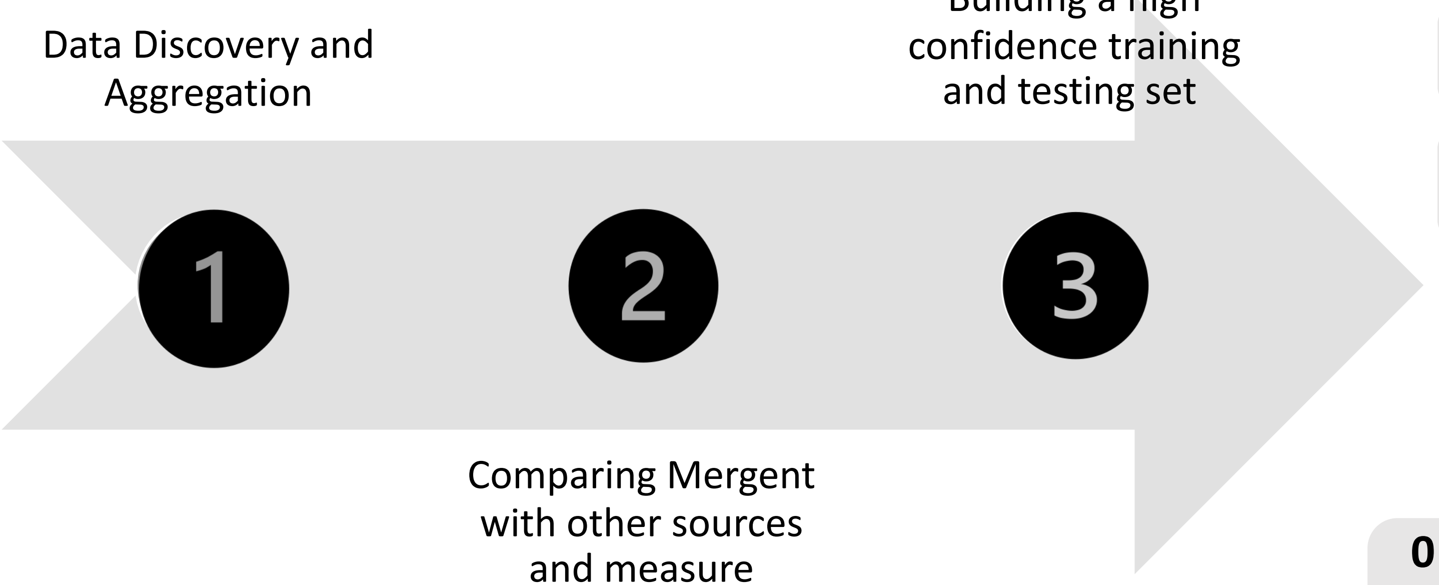
We are **interested in the data provided by Mergent Intellect**, an extensive database containing business information, including:

<u>Industry, founding year, executive names, company location, etc.</u>

However, the U.S Census Annual Business Survey (ABS) reported approximately 38% of minority-owned businesses in 2017, while Mergent Intellect only reported 7% during the same period

Therefore, we have reason to believe that current granular business microdata like **Mergent Intellect underrepresents** businesses owned by minorities

Methodology



1 Data Discovery and Aggregation

2 Comparing Mergent with other sources and measure Mergent's underrepresentation

3 Building a high confidence training and testing set

4 Construct, train and test the classifier:

$$Y = f(X)$$

Input: X = <u>observable predictors</u>

- **Executive names**
- **Language of the Company Name**
- **Location demographics**

Output: Y = <u>Binary Classification</u>

- **0 if the company is not minority owned**
- **1 if the company is minority-owned**

5 Comparing the predicted number of minority owned business with the 2017 Census Annual Business survey.

1. Data Discovery & Aggregation

Mergent Intellect	Number of Businesses
Total Number of Businesses in Fairfax	166K
w/executive names reported	18K
w/executive names reported and flagged as minority-owned	743

Additional data sources to identify minority-owned businesses		
	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

2. Compare Mergent Intellect and other sources

		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

$$\frac{15}{23+15} \cdot 100 = 39.47\,\%$$

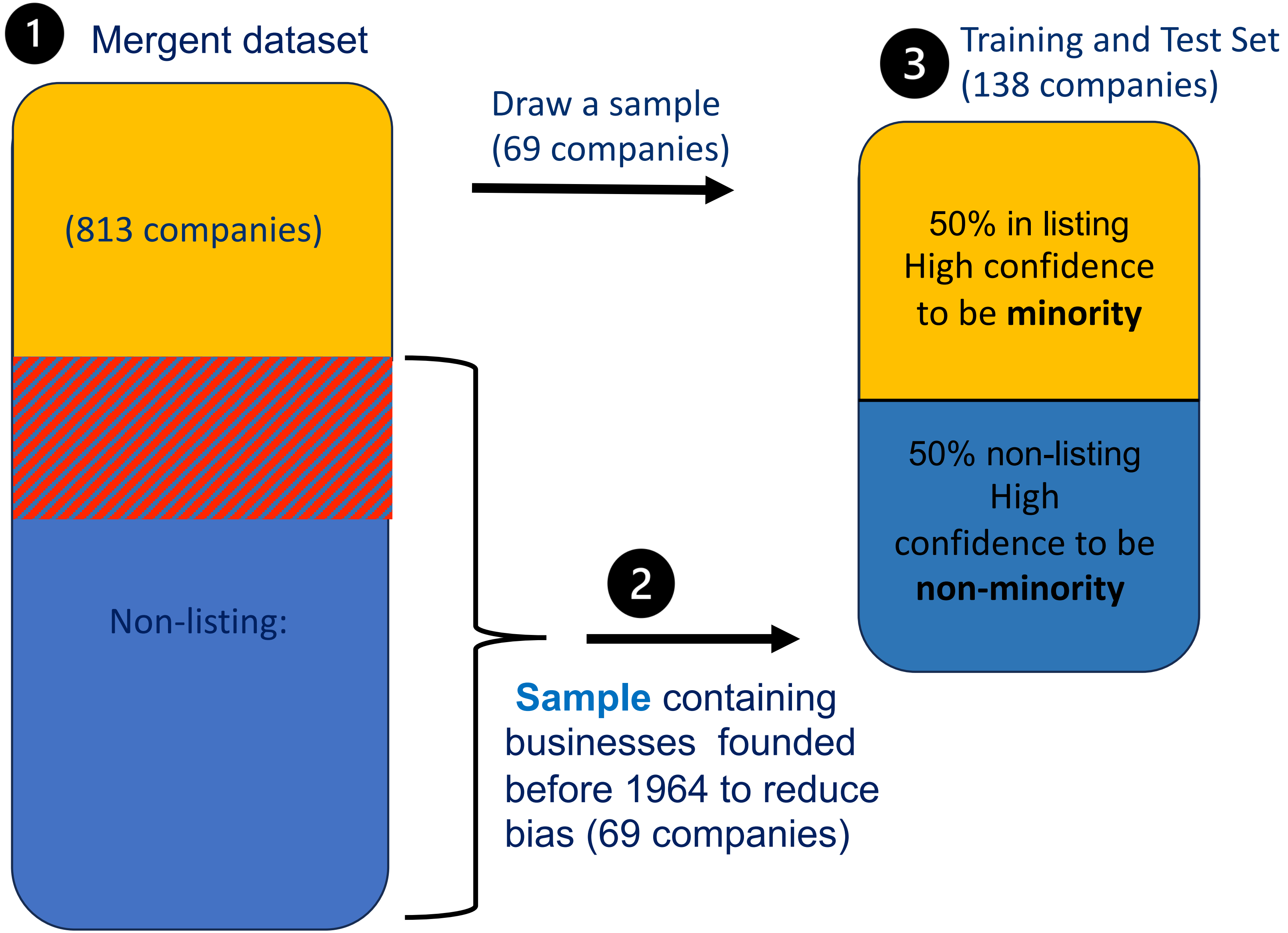
The number of misclassified businesses from Mergent Intellect is **39.47%**. Our goal of our model is to reduce this rate of misclassification

3. Developing a Training and Test Set

Minority owned businesses identify as minority-owned by Mergent Intellect

Minority-owned businesses identify as non-minority owned businesses by Mergent Intellect

Non minority-owned businesses identify as non-minority owned by Mergent Intellect



1 Mergent dataset

(813 companies)

Non-listing:

Draw a sample (69 companies)

2

Sample containing businesses founded before 1964 to reduce bias (69 companies)

3 Training and Test Set (138 companies)

50% in listing High confidence to be **minority**

50% non-listing High confidence to be **non-minority**

4. Constructing the Classification Model

Think of our Model as a Function: Y=F(X), where X represents our observable predictors:

	Predicted Race using Executive Names	Predicted Race using Language of Company Names	Location Demographics
Info:	**Combination of Pre-trained Natural Language Processing (NLP) Models:** • raceBERT • rethnicity	**Combination of Pre-trained NLP Models:** • spaCy • Langdetect	**Based on 5-year Census American Community Survey Data at the Census Tract Level**
Output	Probability that the owner is a member of a racial minority.	Probability that the name of the company is from a predominantly minority-spoken language.	The percentage of minority individuals in each census tract

5. Classifier results and flagging rules

- For each sample of 138 companies, we split it into training (70%) and test (30%) set
- The model was run 10,000 times to produce the evaluation metrics

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%

Flagging rule:

- If a company is identified as minority-owned by Mergent, we report the ownership status from Mergent Intellect
- If a company is listed as non-minority-owned by Mergent, we report the ownership status from the classifier

6. Comparing the Prediction with the Census 2017 Annual Business Survey

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

- The Decision Tree model successfully reduced the misclassification and increased the percentage of minority owned businesses from the Mergent Intellect dataset
- Other models like Probit and Support Vector Machines were also implemented with different configurations

7. Limitations and Next Steps

Limitations:

- There is ambiguity about the data collection process of certain sources
- The filtering mechanism applied to our dataset makes our training and testing sets small

Next Steps:

- Expand our training set by collecting more data and applying relevant data transformations
- Explore how minority-ownership is distributed geographically and across specific industries
- Apply our classifier model to other geographies covered by the Social Impact Data Commons

Questions?

Sources

Yelp: https://www.yelp.com/

Mergent: https://www.mergentintellect.com/index.php/search/index

SBSD: https://sbsd.virginia.gov/

Chamber of Commerce:

- Hispanic: https://www.novahispanicchamber.com/
- Black: https://www.northernvirginiabcc.org/
- Asian: https://www.aabac.org/

Mergent Intellect: https://www.mergentintellect.com/index.php/search/index

AtoZ: https://www.atozdatabases.com/home

Census: https://www.census.gov/quickfacts/fact/table/fairfaxcountyvirginia/POP010220#POP010220

NC Statewide Voter Registration Data: https://www.ncsbe.gov/results-data/voter-registration-data
1964: https://nap.nationalacademies.org/read/9719/chapter/10#191

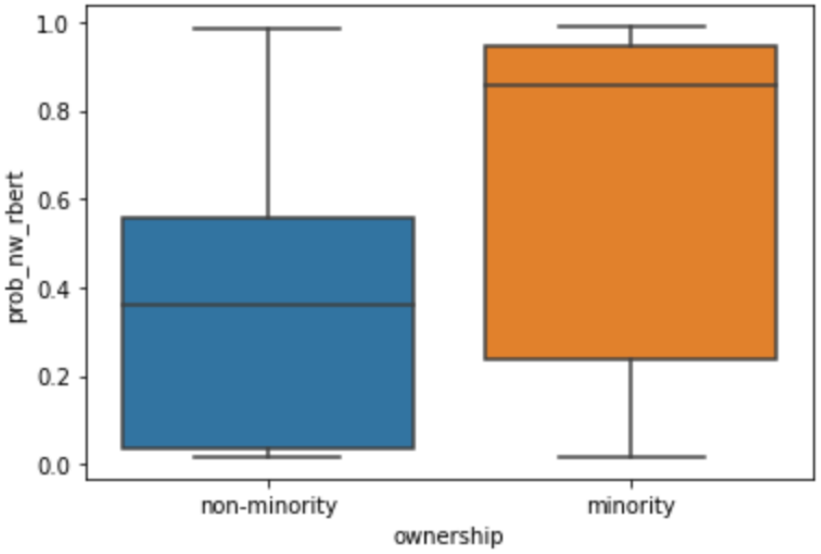
Questions: NLP on Owner's Name

Using the Owner's Name to Predict Minority Business Ownership

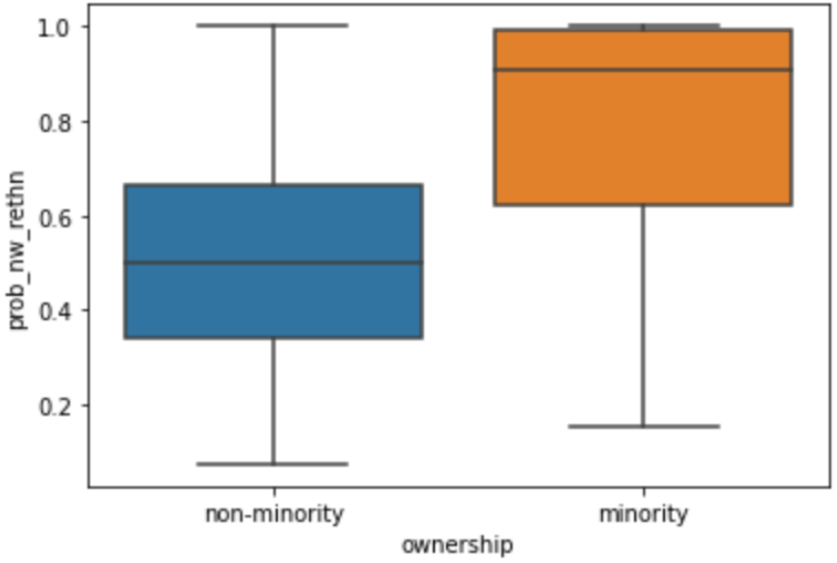
Why we chose to focus on the 'RaceBERT' package and 'rethnicity'?

Boxplots of the average prob to be non-white for raceBert and 'rethnicity'

RaceBERT



Rethnicity



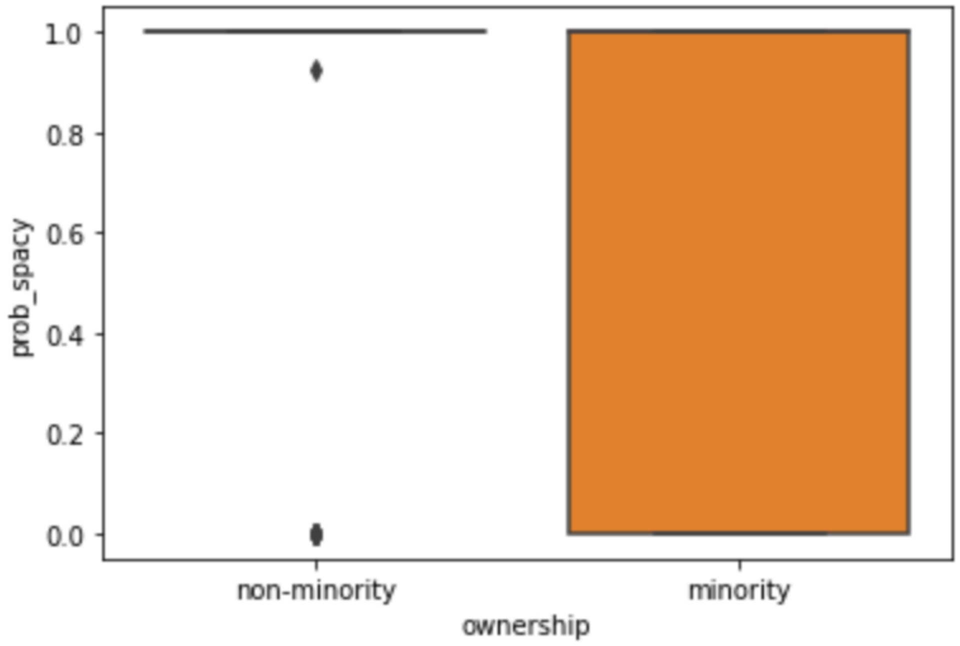
Questions: NLP on Company Name

Using the Company's Name to Predict Minority Business Ownership

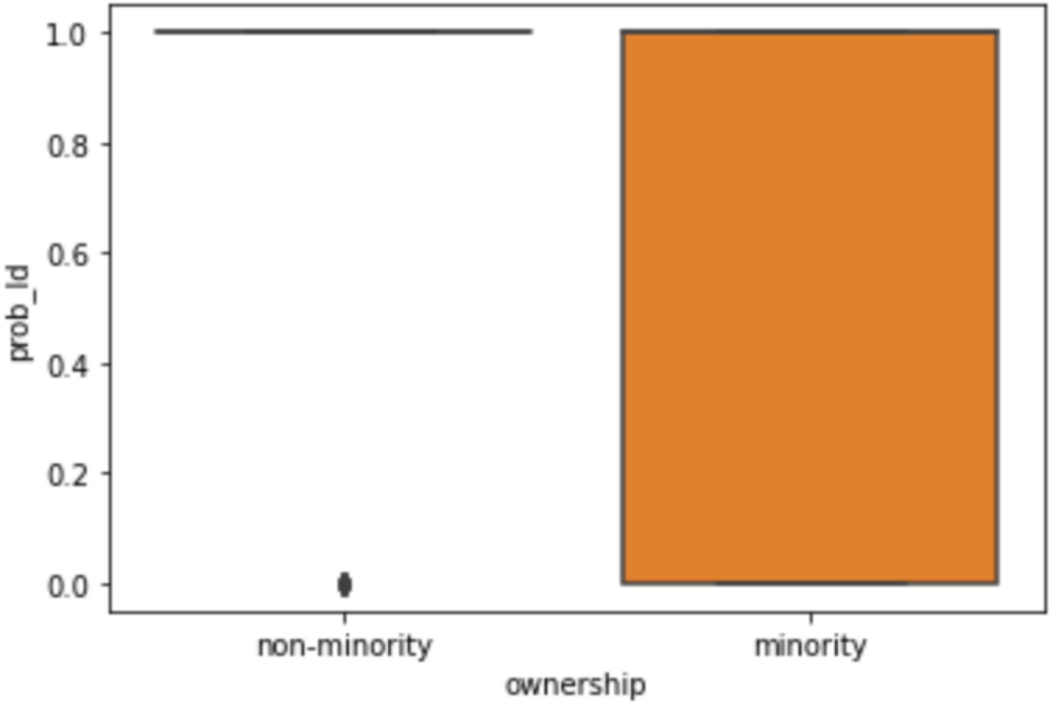
Why we chose to focus on the 'SpacY' package and 'LangDetect'?

Boxplots of the average prob to be non-white for 'SpacY' and 'LangDetect'

SpacY



LangDetect



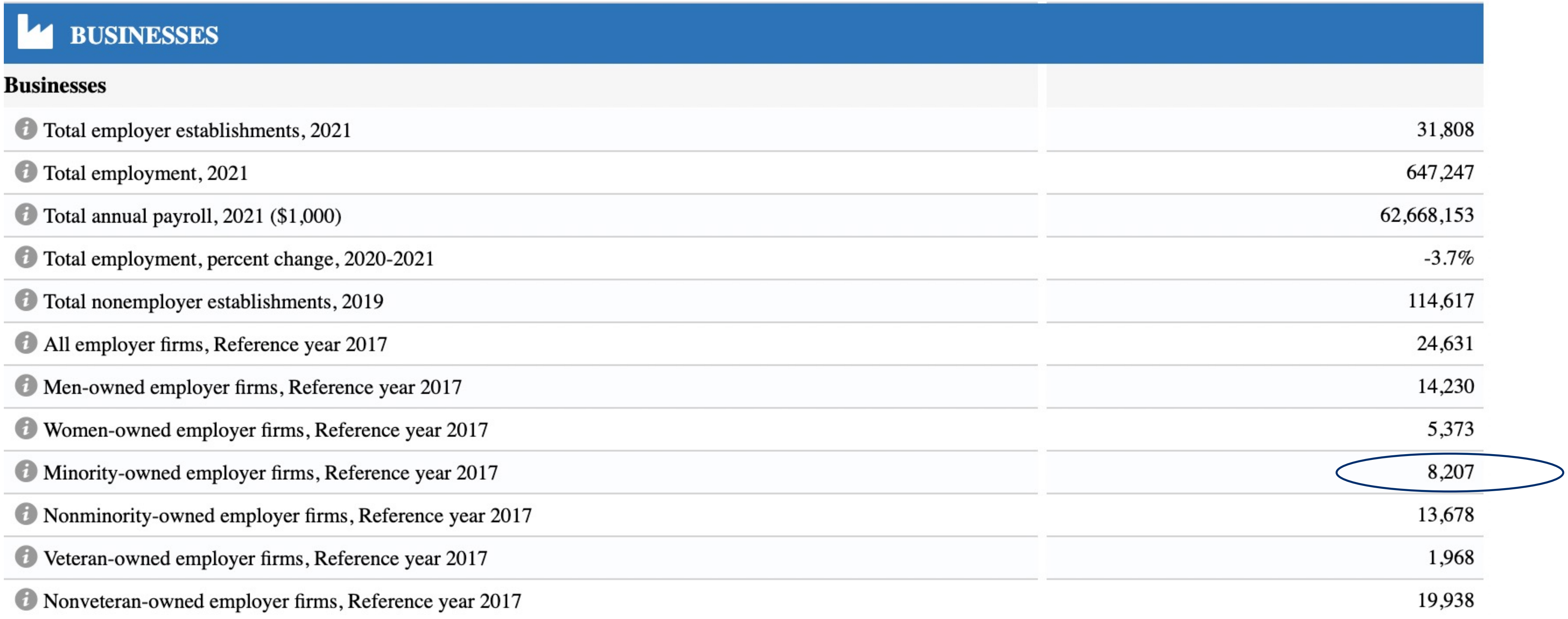
Questions: What is Mergent?

Mergent Intellect is a data aggregator, not a data collector.

Diverse businesses for the most part must self-identify in order to be recognized as a diverse business. Once this info is received through a third-party source, it is maintained through their DUNSRight drivers and Supplier Diversity update processes.

- Database maintains approximately 250 different sources of third-party data
- From there, they narrow down businesses that are 100% sole proprietor ownership and corporations were the ownership of stock is calculated to be least 51 percent. They also eliminate non-profit, publicly held, and government entities.

Questions: US Census data



BUSINESSES

Businesses	
Total employer establishments, 2021	31,808
Total employment, 2021	647,247
Total annual payroll, 2021 ($1,000)	62,668,153
Total employment, percent change, 2020-2021	-3.7%
Total nonemployer establishments, 2019	114,617
All employer firms, Reference year 2017	24,631
Men-owned employer firms, Reference year 2017	14,230
Women-owned employer firms, Reference year 2017	5,373
Minority-owned employer firms, Reference year 2017	8,207
Nonminority-owned employer firms, Reference year 2017	13,678
Veteran-owned employer firms, Reference year 2017	1,968
Nonveteran-owned employer firms, Reference year 2017	19,938

The last record for minority ownership was in 2017

Questions: Ethics

Doesn’t this model seem a little unethical?

We understand how a binary racial classification model might seem unethical in most cases. However, we will prevent this by…

1. Aggregating our qualitative results into quantitative results (number of businesses in the listing (labelled as minority) and number of businesses who are not (not labelled as minority)

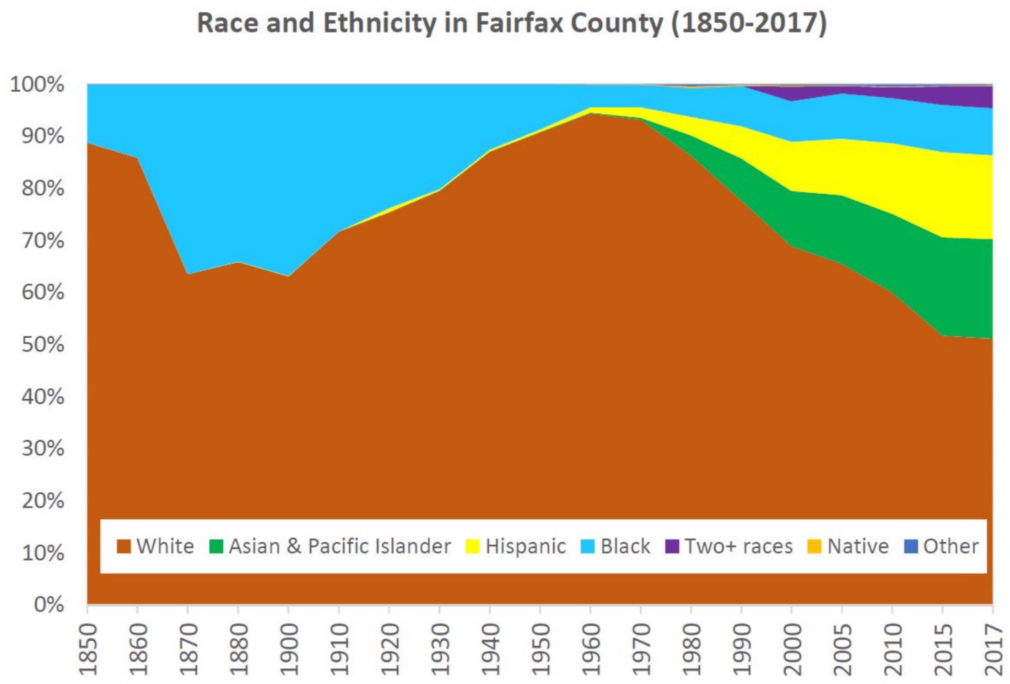
2. Inhibiting public access to our physical predictive model to prevent its misuse

3. Binary Classification result instead of Individual Race Classification

Questions: Why 1964?

Census data indicates that the population of Fairfax County in 1964 was 90% non-minority individuals

Also, the civil rights act was enacted in 1964, and that event greatly bolstered minority business ownership



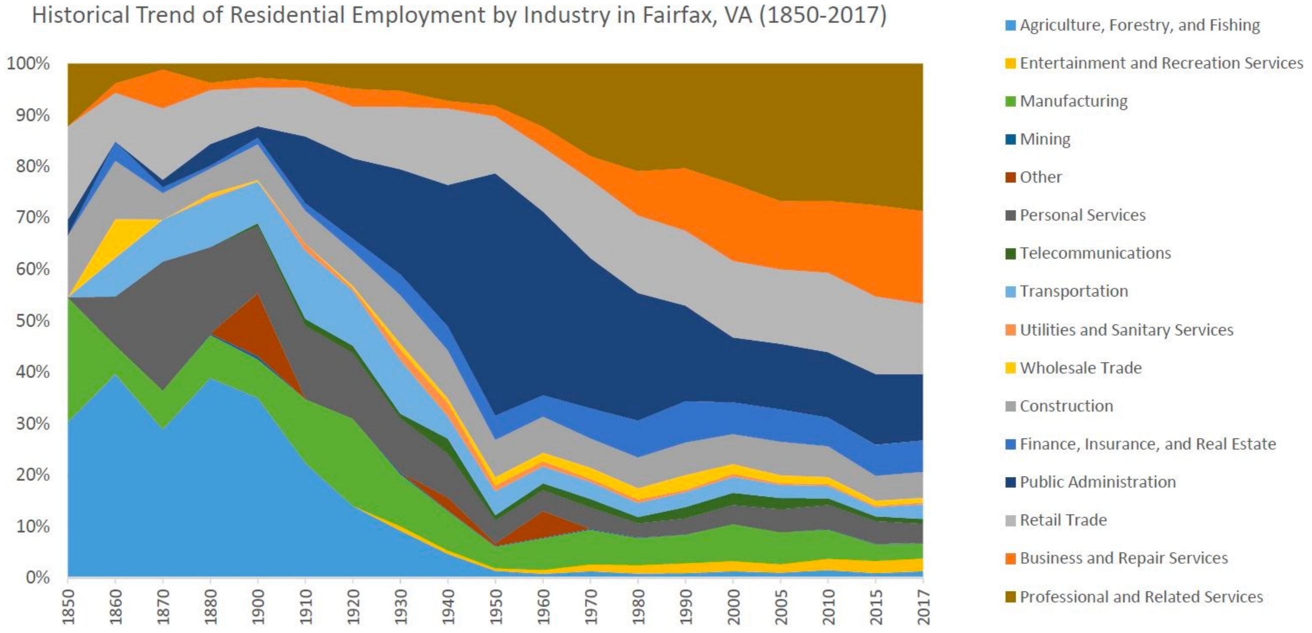
Note: The race categories don't include hispanics. Hispanic includes population of any race.

Source: U.S. Bureau of Census, Decennial Census (1850-2010), American Community Survey (2005, 2015 and 2017).

Source: https://storymaps.arcgis.com/stories/f74a8fbad837435b8e901cc9c04aa345

Questions: Examining industry in the future

We hope to compare the distribution of industries in Fairfax County to the proportion of minorities in each industry to help understand the trends that may be occurring

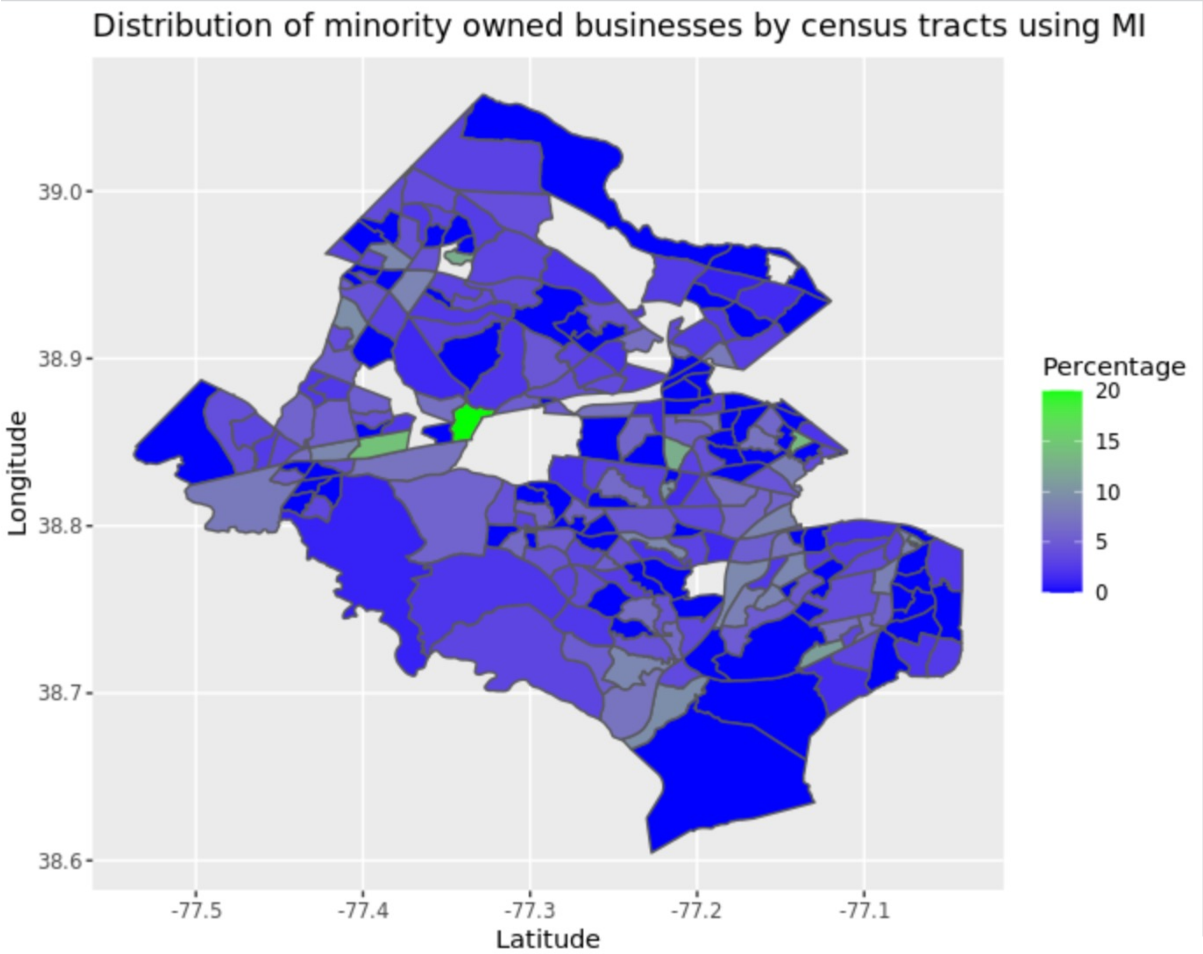


Note: Residential employment refers to the number of Fairfax County residents, regardless of their place of employment.

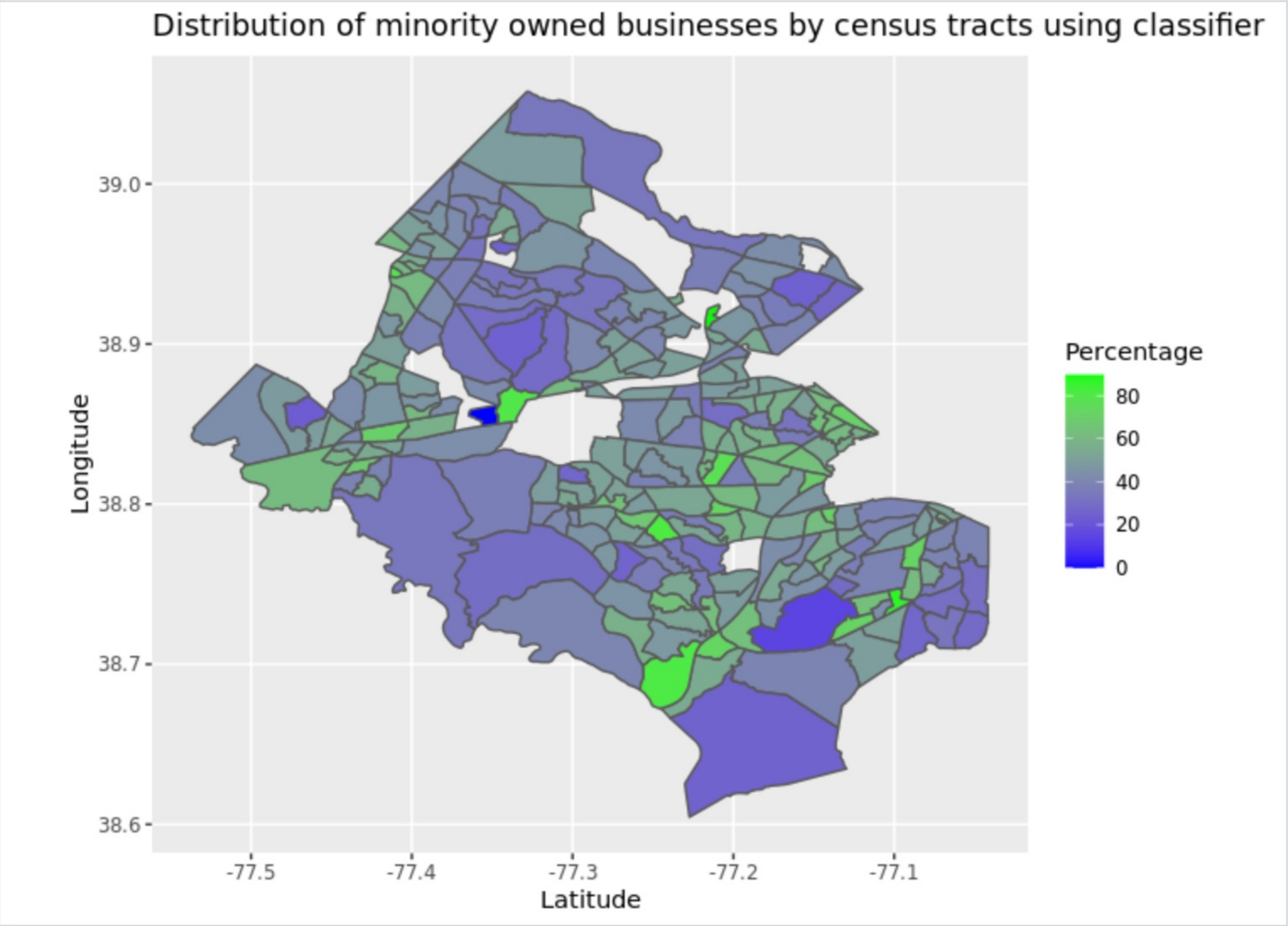
Source: U.S. Bureau of Census, Decennial Census (1850-2010), American Community Survey (2005, 2015 and 2017).

Source: https://storymaps.arcgis.com/stories/f74a8fbad837435b8e901cc9c04aa345

6. GIS map



Distribution of minority owned businesses by census tracts using MI



Distribution of minority owned businesses by census tracts using classifier

The map on the left is the percentage of minority owned businesses in each census tract before the model was applied, and the map on the right is after the model was applied.

This was applied to the 18k data points from Mergent Intellect of companies with executive's names reported