Project Technical Report

Pushing P Breakers

Group 1

Parth, Ahad, Luis, Sharif, Marjea

New Jersey's Uninsured

Introduction

The economic downturn caused by the coronavirus pandemic has renewed attention on health insurance coverage as millions have lost their jobs and potentially their health coverage. The Affordable Care Act (ACA) sought to address the gaps in our healthcare system that left millions of people without health insurance by extending Medicaid coverage to many low-income individuals and providing subsidies for marketplace coverage for individuals below 400% of poverty. Following the ACA, uninsured non-elderly Americans declined by 20 million, dropping to a historic low in 2016. However, beginning in 2017, the number of uninsured non-elderly Americans increased for three straight years, growing by 2.2 million from 26.7 million in 2016 to 28.9 million in 2019. The uninsured rate increased from 10.0% in 2016 to 10.9% in 2019.

With the rising uninsured population, it becomes essential to keep track of those changes and be able to map a specific number of uninsured individuals to a specific geographic area. This importance is dictated by the need for private insurance entities to target particular groups for their market segmentation and for public insurance entities to focus their programs and policies that aim to reduce and eliminate the number of uninsured in the target area. While that type of information is available on the web and other resources, it may not necessarily be well organized, compact, and readily available to be consumed by those entities. Therefore, the ultimate **objective** of this research work is to locate resources related to the uninsured population, sort and organize that information into a database, and analyze and provide predictions regarding the uninsured population within a specific geographic area, given a set of

demographic characteristics. There may also be recommendations regarding what areas and groups those entities need to zero in on.

Exploratory Questions

To conduct this type of research, we developed some questions and hypotheses that guided us through our work, helping us to stay focused on our subject:

1. Which city has the highest uninsurance rate?
2. How many counties have a population of uninsured people of 8% or more?
3. How does income change the amount of people being uninsured?
4. Is there a race that has a higher uninsured population?
5. Which sex has a higher population of uninsured people?
6. What age range has the most uninsured people?
7. Does employment affect the amount of the uninsured population?
8. Hypothesis test: the larger the population, the higher the uninsured rate.
9. What area and demographic group can Prudential target for insurance sales in NJ?

Dataset Introduction

We used several resources and datasets in this research, including census data. Below is the list of those datasets:

1. Small Area Health Insurance Estimates 2019 (SAHIE)
    a. This dataset contains demographic information about health insurance coverage and demographics in the counties by a single year in the United States.
2. NJ Uninsured
    a. Contains data about the uninsured population in New Jersey broken down by age, race, and sex down to the county and city location
3. NJ Unemployed
    a. This dataset has information about the unemployment rate in each county and city location in New Jersey
4. NJ Income

a. Shows the median household income in every county and city in New Jersey

5. Cartographic Boundary Files
    a. Contains files to show county boundaries for selected geographic areas
6. Coverage for the Household Population by States
    a. Shows the populations for the state of New Jersey
7. Unemployment by Counties
    a. Gives a table that shows the characteristics of the unemployment population in each county in New Jersey
8. Census Tract in NJ
    a. Shows the boundaries for the county borders in New Jersey

Research Process and Discoveries

We downloaded all datasets as CSV files and stored them in our database storage account. After conducting an exploration of our datasets, we did some ETL to get all necessary variables into one SQL database using Azure databricks and data factory and set up a pipeline. We also used ML algorithms to conduct our predictions and presented our findings through visualizations using the Dash platform.

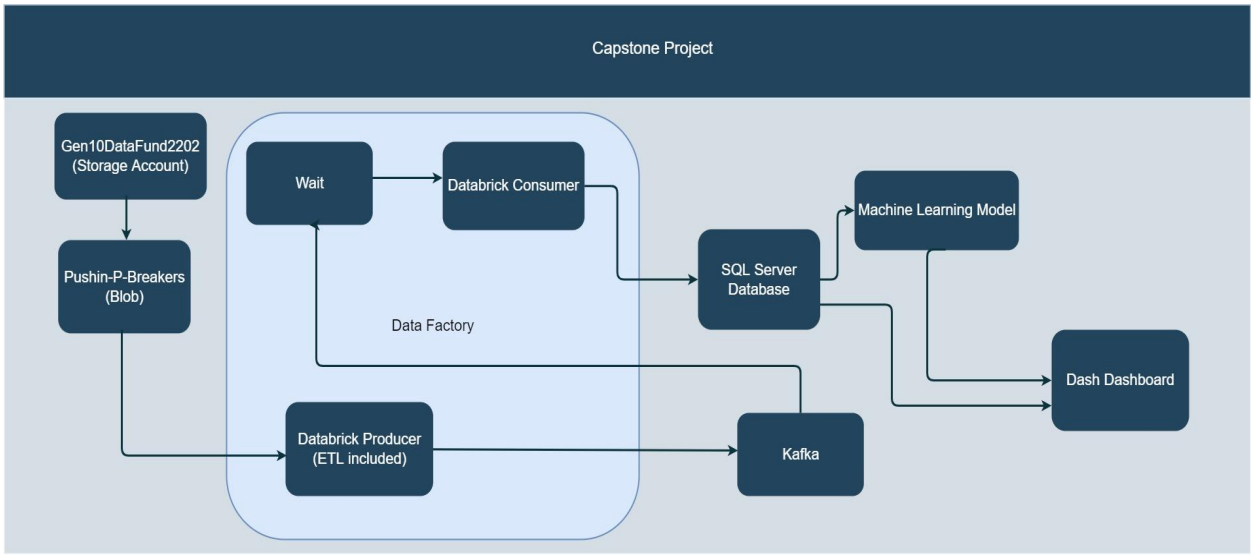

Figure 1. Data Platform

After having all our necessary data ready in our database we did the following analysis and discovered interesting results:

Regarding Q1.

One of the first criteria that assist in understanding the problem of uninsurance is to identify the geography of uninsured residents on several levels. We start by locating the top 10 cities with the highest number of uninsured residents:

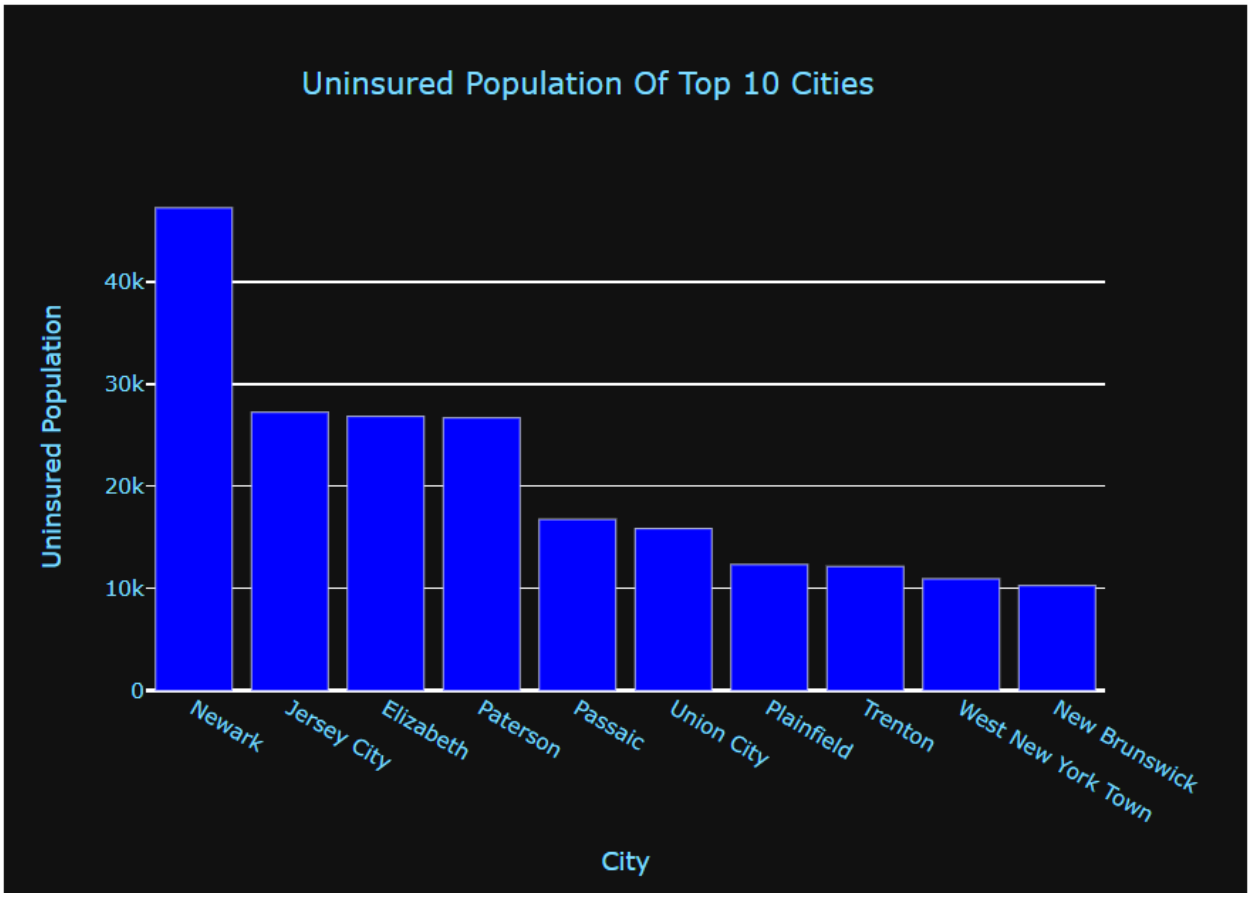

Figure 2. Top 10 cities of NJ with the highest number of uninsured residents

As seen from *Figure 2,* almost all cities with the highest number of uninsured people are located within the north-eastern counties of NJ, with Newark having almost double the number of uninsured residents compared to its closest peer cities. This can be explained by two factors: a) most of those cities have the closest proximity to New York City, and b) because of the proximity to New York City, they are among the highest-density cities in NJ. We can safely make

our first observation which is that the number of uninsured residents of an area is impacted by how dense the population is within that area.

Another geographic level to look into is counties. During our research, we discovered that the national average rate for those uninsured is 8%. Thus, we wanted to identify counties that fall above the national average:

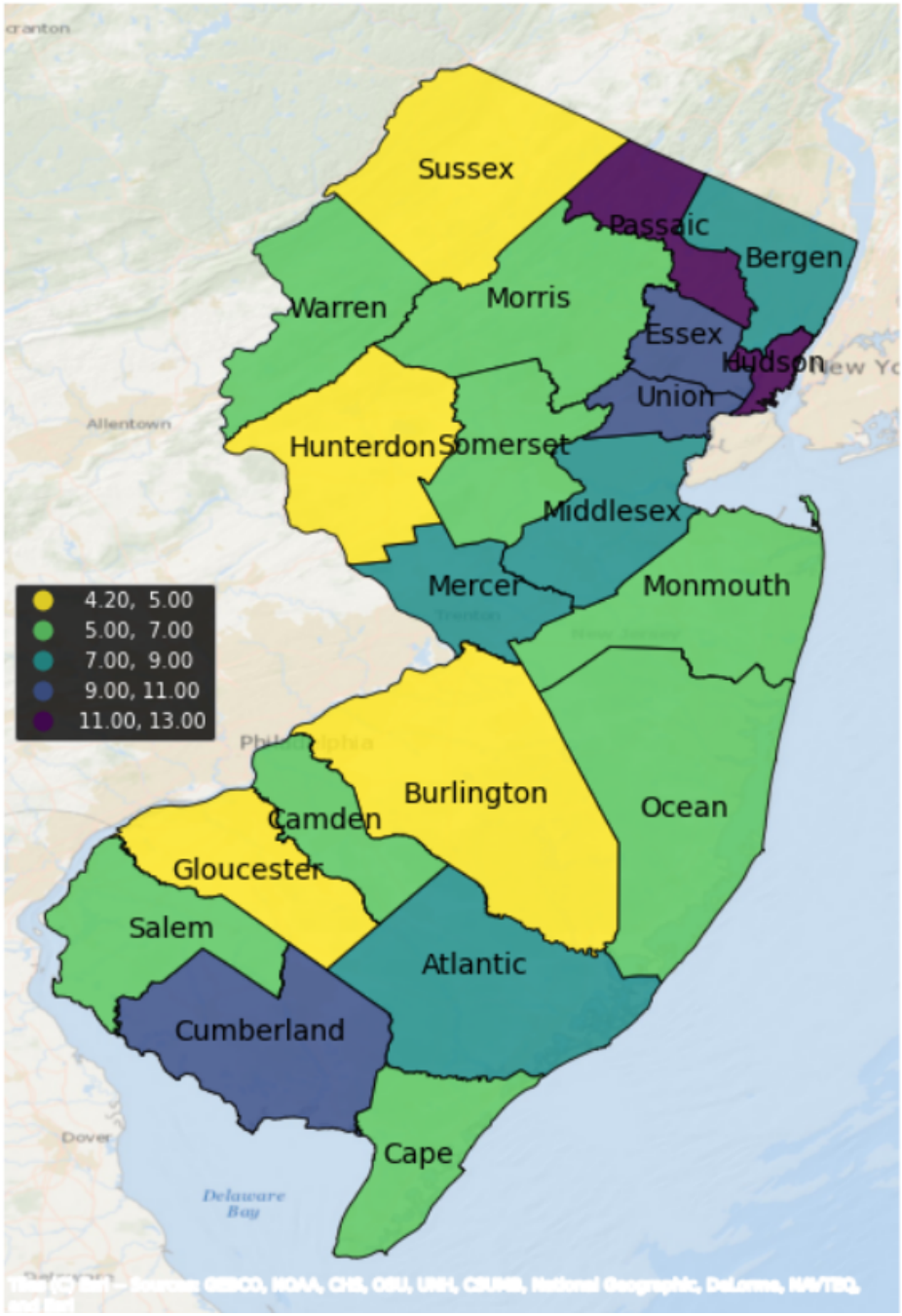

Figure 3. Counties of NJ broken by uninsurance rate

The first observation in *Figure 3* is that counties that are closer to the center of metro areas have the highest rate of uninsured residents. Primarily, northeastern counties like Hudson, Passaic, Essex, and Union are among the top 5 counties with the highest y insurance rate, and at the same time, the closest counties to the centers of the metro area. On the other side of the map, in the south and south-eastern part of the state, we see the counties Cumberland and Atlantic have uninsurance rates above the national average. These counties are also located closest to the center of the metro area, in this case, Philadelphia. These counties, primarily those located in the northeastern part of NJ are some of the densest counties in the state. This observation strongly supports our findings in question 1.

Another criterion we decided to look into is income. The rational assumption would have it that the higher the income level of a geographic area, the lower the uninsurance rate, as residents with better income and jobs would be able to get coverage through their employer's plan or buy one on their own. To check this relationship we created the following figure:

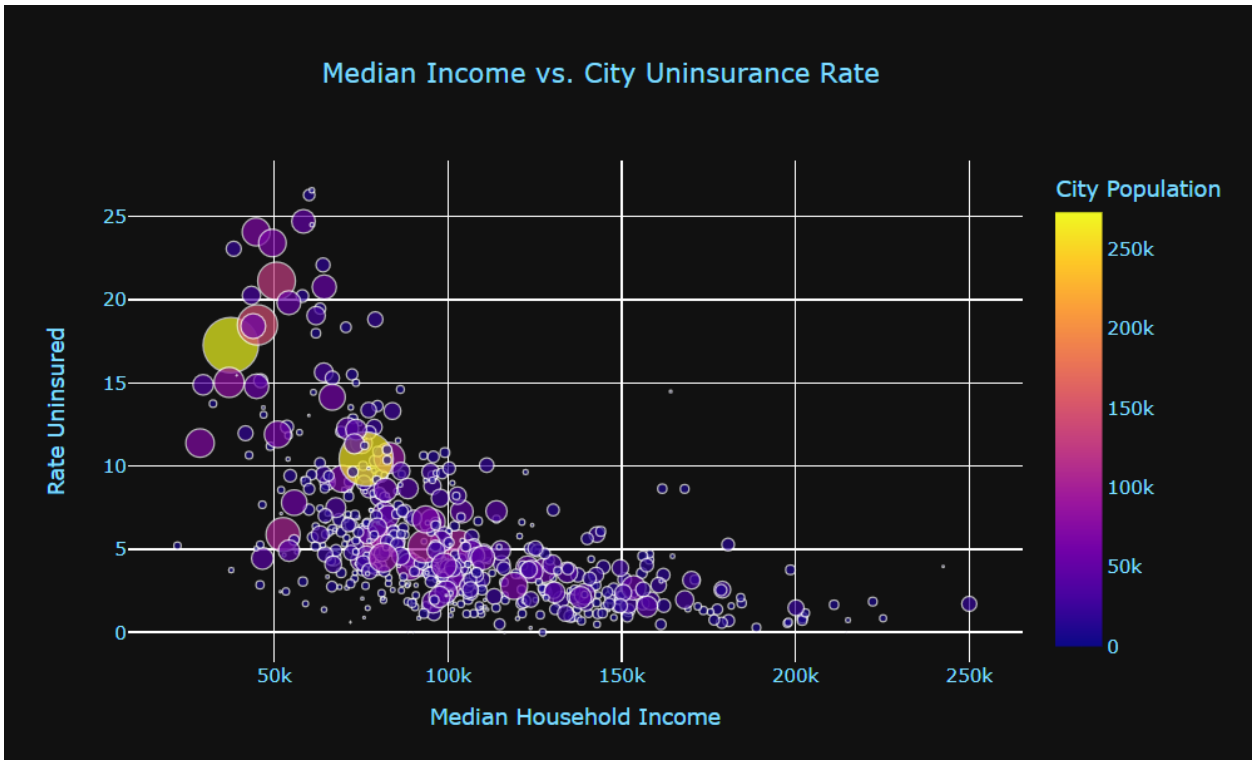

Figure 4. Scatter plot of the relationship between income and uninsurance rate of different cities (city population sizes are provided for reference to previous questions)

The first observation from *Figure 4* is that there is a negative correlation between income and the uninsurance rate of cities. The higher the income the lower the rate tends to be, which to a degree confirms the assumption we had. However, this relationship only proves itself with income starting at around $40,000. How about residents with incomes lower than $40,000? Well, as we all know, families and individuals who live below the federal poverty line can count on state coverage. This is why families and individual residents of cities that fall within those federal poverty lines do not really show as uninsured in *Figure 4*. Lastly, it seems higher populated cities tend to have higher uninsurance rates. This also can be explained by the geographically close proximity of those cities to centers of metro areas and their population density.

Among the demographic characteristics of the uninsured population in NJ, race, age, and gender stand out. Are the uninsured groups evenly distributed among the population or is there a significant difference? If so, what causes it? We start by looking into the race category:

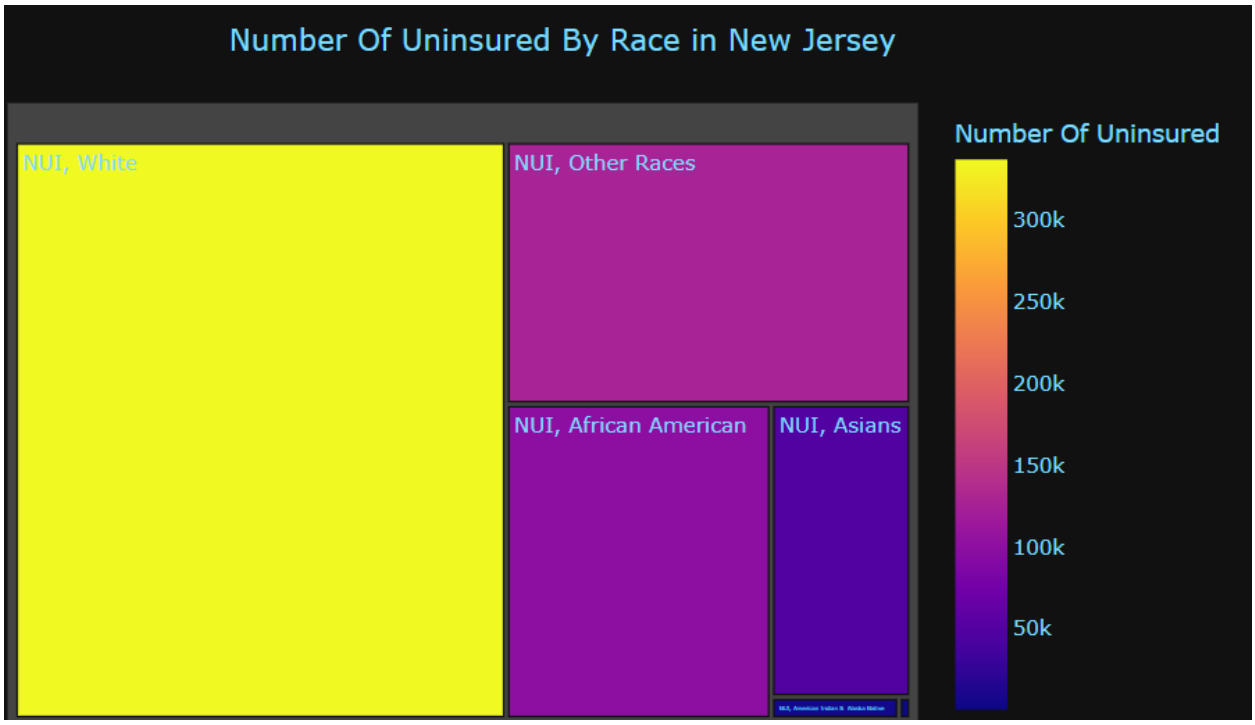

Figure 5. Uninsured population by race

After a quick glimpse over *Figure 5,* we can conclude that the majority of the uninsured population in NJ is white. However, this type of distribution is not primarily dependent on the

uninsurance factor, but rather on the race distribution of the total state population. In NJ, over 70% of the population is white, followed by about 15% being African-American, 10% Asian and the remaining being a mix of other races. *Figure 5* roughly represents the same distribution.

Another characteristic is uninsured population distribution by gender:

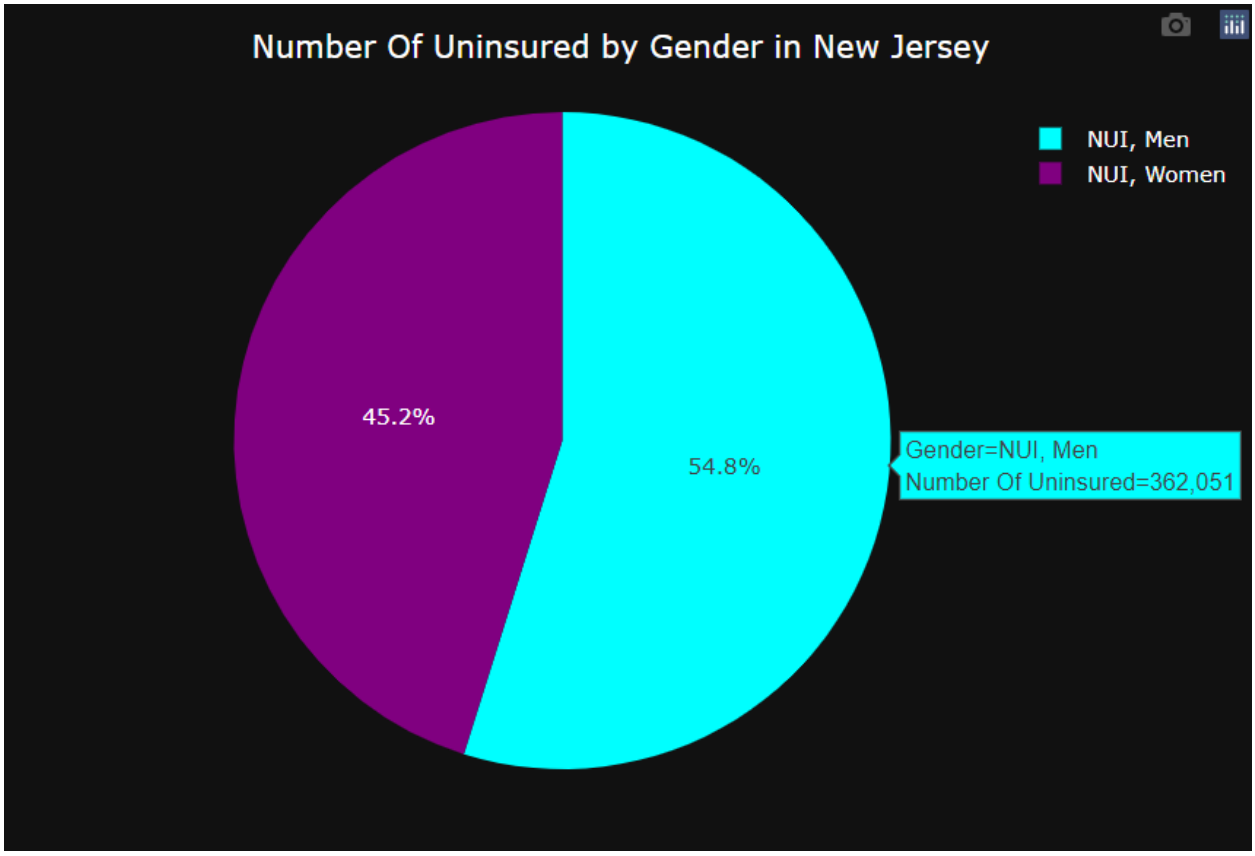

Figure 6. Uninsured population by gender

While the total population of females in NJ is higher than males, here we realize that within the uninsured population the number of males that are uninsured is slightly higher than females. One of the reasons the results are as such is that in NJ, female adults have higher chances of getting state coverage compared to same-age male adults. This is primarily because of pregnancy conditions. Another reason is that some families with tight incomes may choose to have only a part of the family, predominantly females, have coverage, leaving male members with little to no coverage.

Age is also another important characteristic to look into, since the uninsured population may be impacted by age and health conditions age ranges may cause:

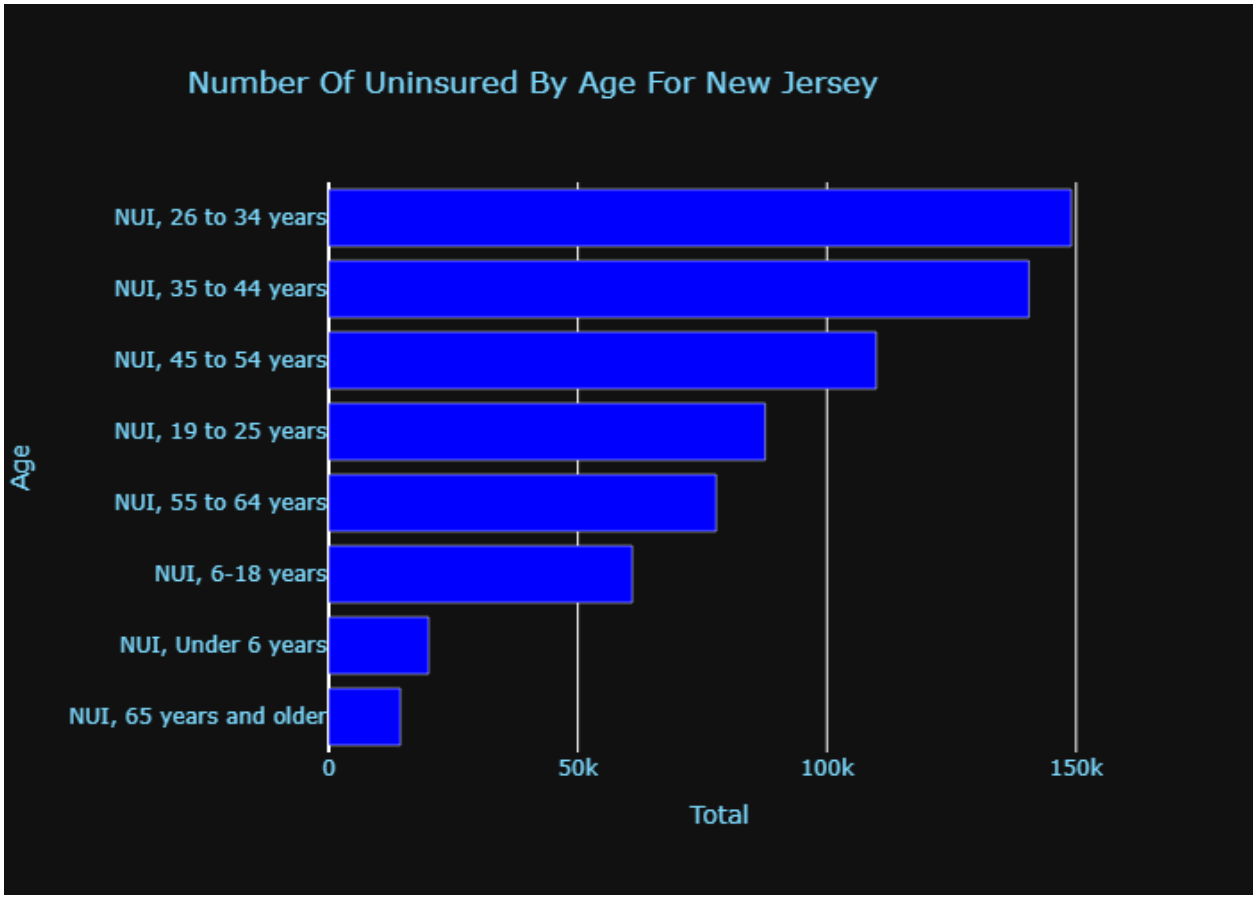

Figure 7. Uninsured population by age group

Population groups between the ages of 26-54 are more likely to be uninsured compared to the other age groups. When individuals are young, they are likely to be covered under their parents' coverage or state coverage. Once they reach 26, they are no longer eligible to be dependent on parent coverage, nor are they usually eligible for state coverage. This is a period when many face the challenge of affording coverage and there are not many options for low-income residents. Once they reach the age of 55 and over, they become Medicare eligible, which by this age, increases the probability of an individual being insured.

Lastly, we considered the employment state of the uninsured population. Our rational assumption was that within the uninsured population, the majority would be unemployed. To check this hypothesis we created *Figure 8*:

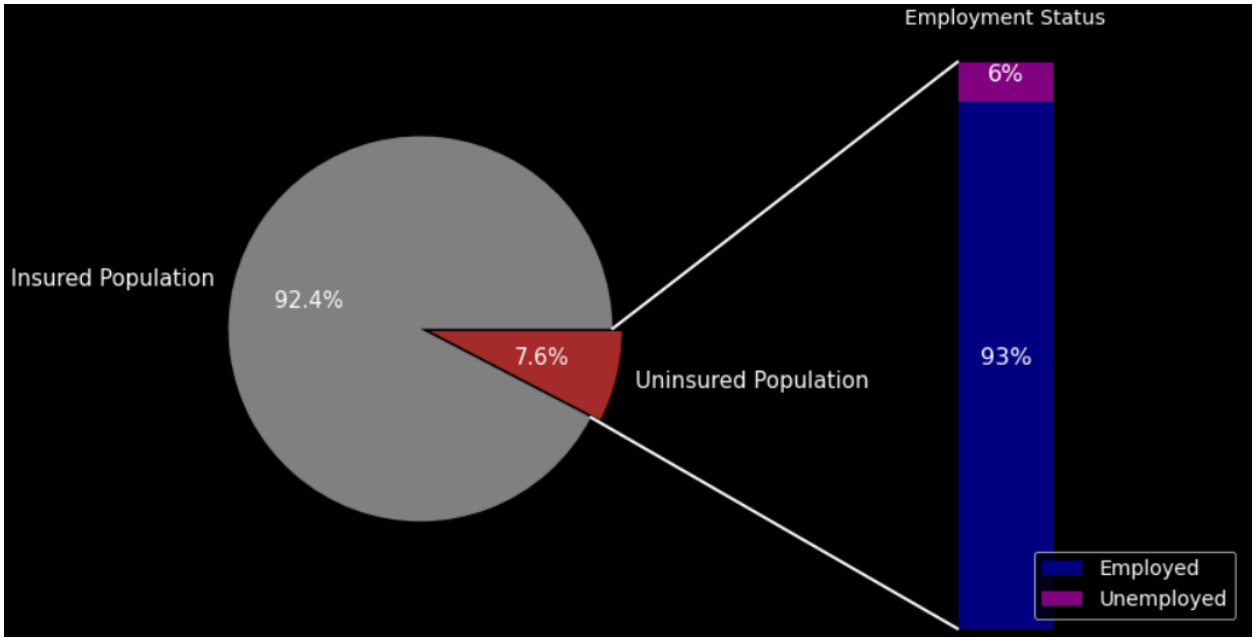

Figure 8. Uninsured population by employment status.

As seen in *Figure 8* within the uninsured population over 90% are employed individuals, and only 6% are unemployed. This is very interesting, not only because it rejects our rational assumption, but it gives insight into how employer plans lack affordability for even their own employees. The employed category also represents business owners and employees of small companies that don't offer coverage at all. However, the majority represent employees of companies that do offer some type of coverage.

After our research and analysis, we built a list of evidence-based findings from the uninsured population in NJ:

1. **Close proximity to the centers of metro areas (location):** The closer the location to the centers of metro areas, the higher the number of uninsured.
2. **Population density:** The higher the population density the higher number of uninsured.

3. **Income level:** Generally, higher income leads to lower uninsurance rates.
4. **Race:** The majority of the uninsured population is white, but that reflects the proportion of the white population from the total state population.
5. **Gender:** Males represent a slightly higher uninsured population than females.
6. **Age:** Those within the age range of 26-54 are at more risk of being uninsured compared to younger or older age ranges.
7. **Employment status:** Employment status doesn't seem to have a significant influence on the number of uninsured, where over 93% are employed.

So, who is a typical uninsured person in New Jersey? **This is an individual that tends to be a white male 26-54 years old who is employed and with an income of around $40,000.**

What area do uninsured individuals tend to live in? **Close to the centers of metro areas and/or urban areas, and within high-density populations.**

Exploratory Data Analysis

Our goal is to use existing information about uninsured populations at the city level to predict the uninsured population information at the census tract level. This type of out-of-sample prediction is known as **interpolation**. Typical interpolation methods include Kriging (Gaussian Process), Inverse Distance Weighting, Spline, and Linear Interpolation. However, these methods make the assumption that the locations of interest can be modeled as points. However, geographics (counties, cities, census tracts, etc.) are not point-based values but rather areas on a 2-D map. Our interpolation model must reflect that our geographical locations are polygonal objects and estimate our variables accordingly.

Machine learning Algorithm

With the stipulation that we are modeling our geographics as polygons in mind, we are using the following Machine Learning Algorithm to predict the goal:

Spatial Area Interpolation

Spatial Area interpolation is a technique for estimating values contained within target geographies with unknown values.

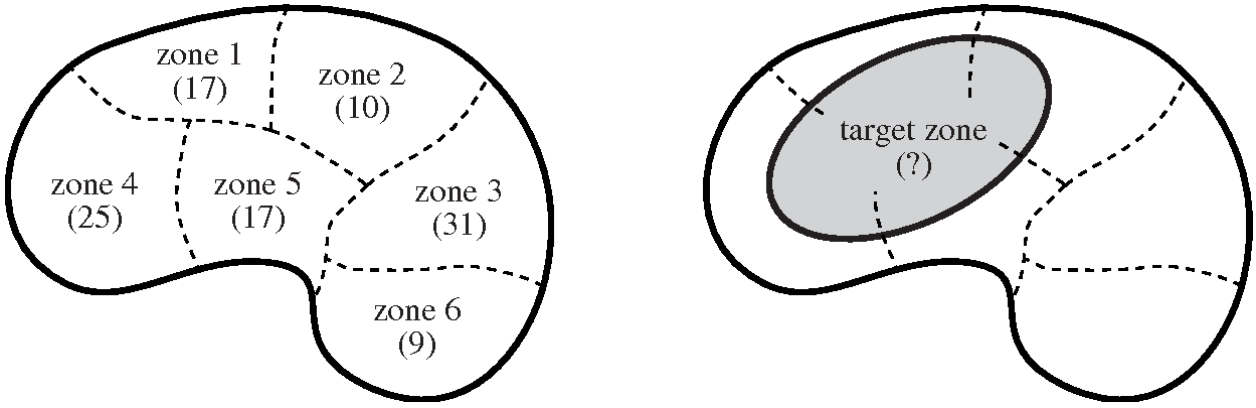

Figure 9. Spatial Interpolation Example

The process works by first estimating the overlap a target geometry (with unknown values) has with neighboring source polygons (with known values). Variables are weighted based on this overlap and then reaggregated to match the target polygon geometries. In addition to overlap, variables are also estimated based on whether they are **intensive** properties or **extensive** properties:

- Intensive Property: independent of the size of the system (Population Density, Concentration, Melting Point, etc.)
- Extensive Property: dependent on the size of the system (Population Count, Mass, Volume, etc.)

In the case of our dataset, “Unemployment Rate (16 and Over)” is the only *extensive* property. All of the other variables are counts of uninsured by demographic characteristics, which are *intensive*.

After building the model, we tried to predict the uninsurance rate of NJ census geographical tract areas and compared our results with the actual rates:

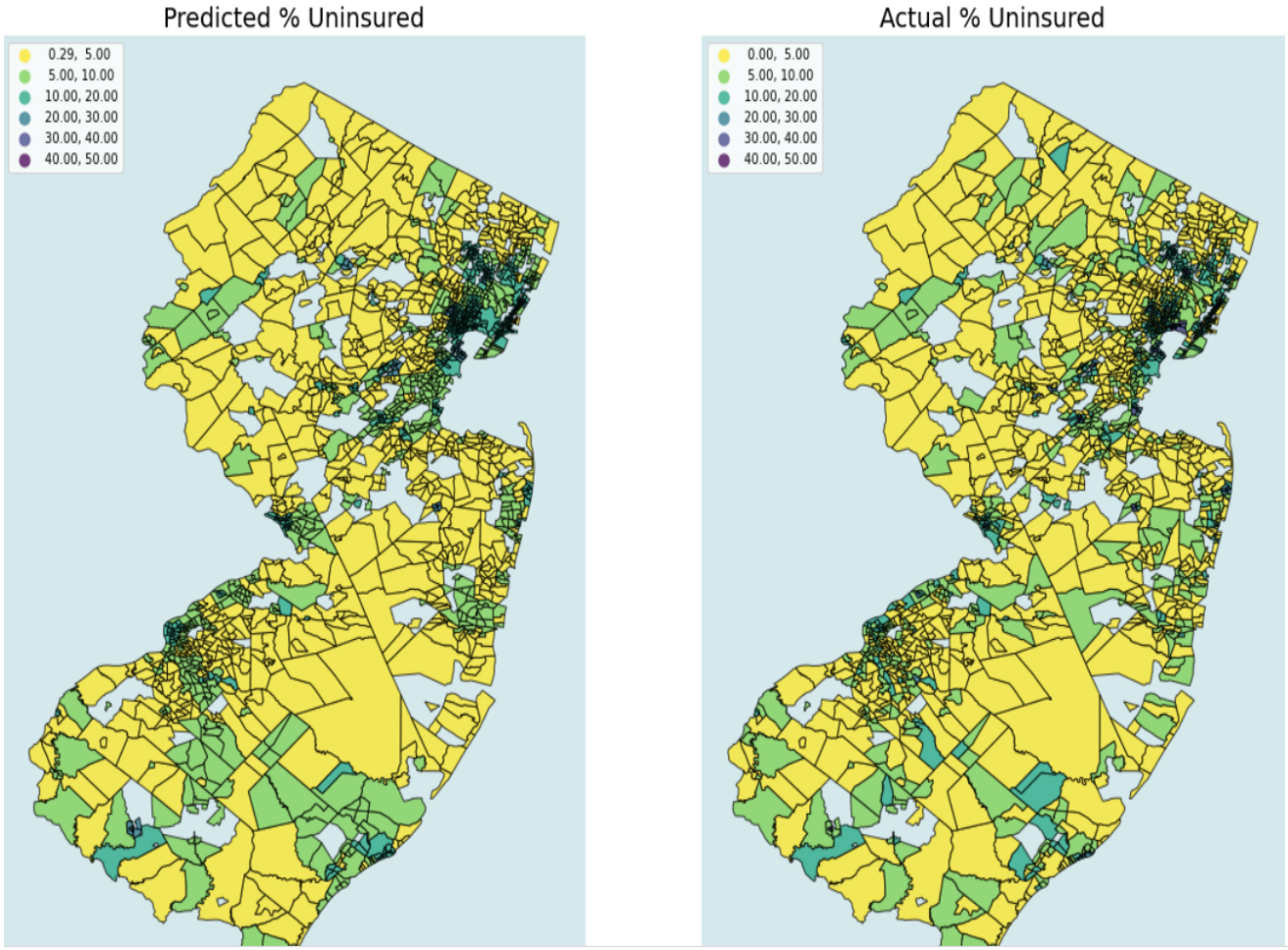

Figure 10. Actual uninsurance rate vs. predicted (NJ census tracts)

As seen in *Figure 10* our model was able to predict the uninsurance rate of given geographic areas (tracts in this case) with a high level of accuracy. To assess the model’s

accuracy for data with a wide range of values, we next plotted the uninsurance rate at the census tract level for the white population of New Jersey. The results are below:

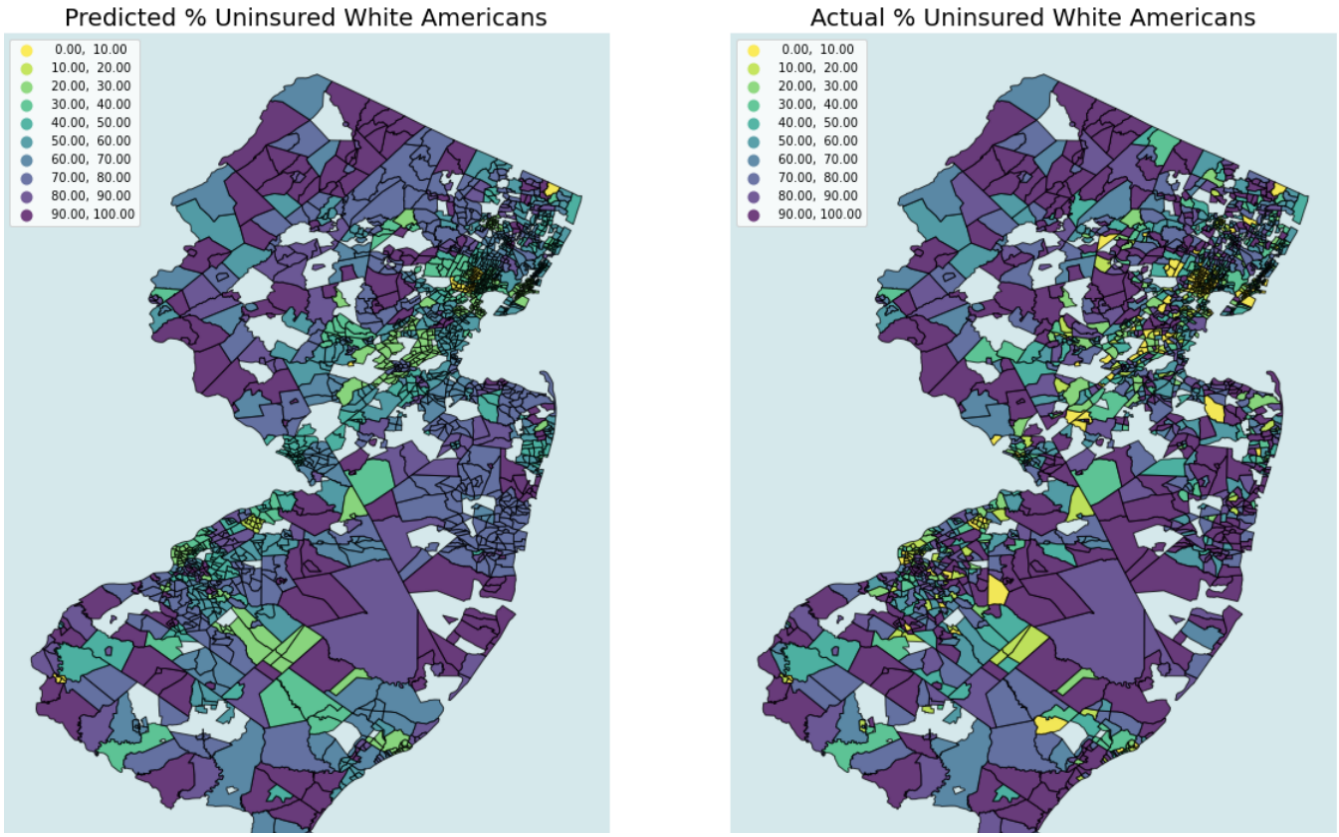

Figure 12. Actual uninsurance rate vs. predicted within NJ census tracts for White Population

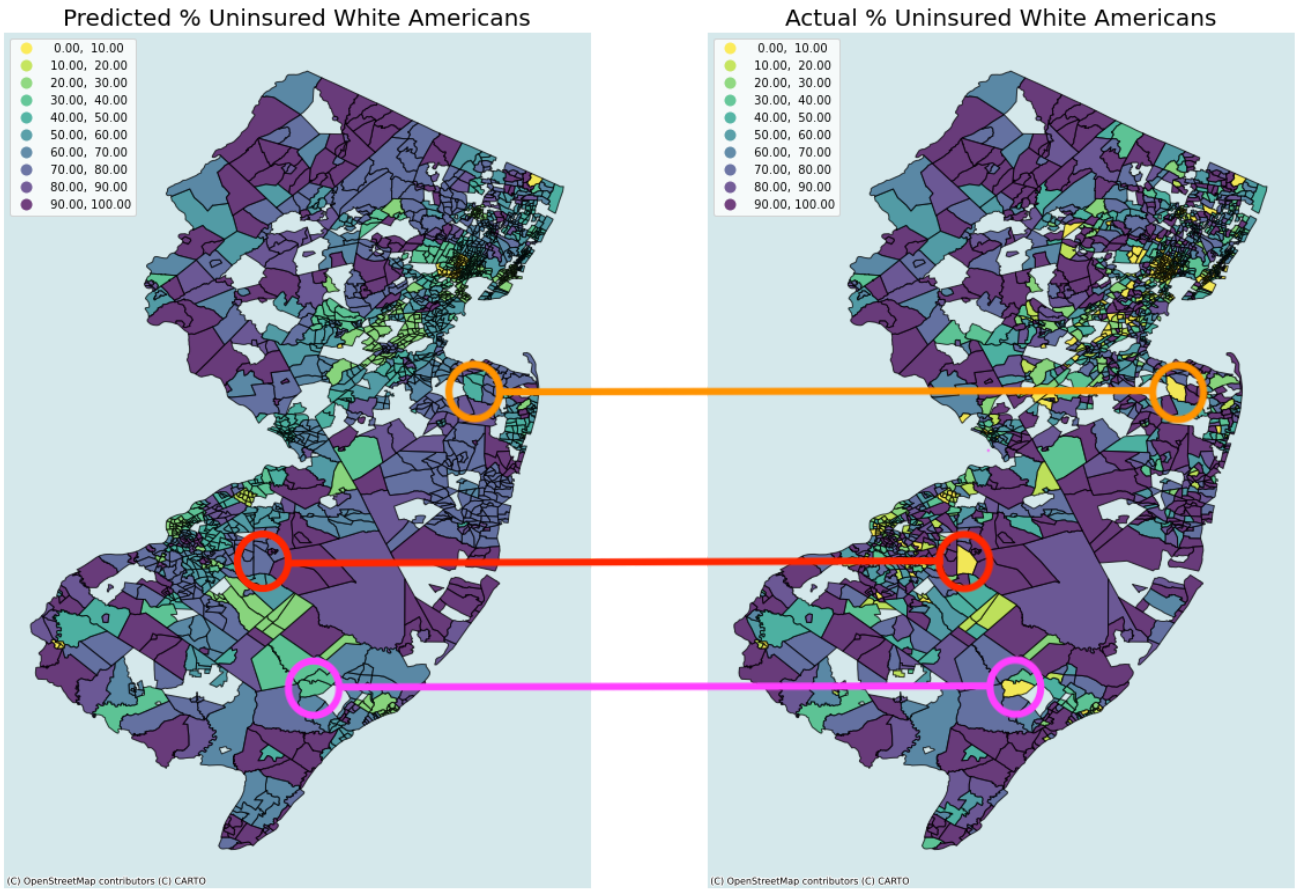

Figure 13. Actual uninsurance rate vs. predicted within NJ census tracts for White Population (Outliers circled)

It becomes apparent that the model is susceptible to outliers. This is best seen in Figure 13, where some of the local outliers are circled. Because these local outliers are surrounded by cities with a high uninsurance rate, the model subsequently predicts a high rate of uninsurance for the local outlier. This is part of a much larger issue known as the **Modifiable Areal Unit Problem** (MAUP). In short, by aggregating point-based values (such as population counts and density) into a 2-D polygon, we are introducing statistical bias into our model that pushes the model's prediction in a certain direction. Put another way, data tabulated for different spatial scale levels or according to different zonal systems for the same region will not provide consistent analysis results. For example, for our model, we chose data at the city level to predict

data at the census tract level. If we had chosen US census data at the neighborhood level of a city—which does not exist but for the sake of argument, assume it does—and then used *that* to predict data at the census tract level, we would get *dramatically* different results. This means there is no consistency in the results of our model across different spatial reference frames. All of this, again, is a consequence of using aggregated data from the census (a singular data point) and extending that to a 2-D area. We discuss solutions to this problem (and others like it) in the section on improvements to the model.

Advantages/Disadvantages of Spatial Area Interpolation Model

Advantages:

- Highly extensible
    - This model can easily be built upon with new data related to the distribution of uninsured persons in New Jersey
- Variable independence
    - Since all of our variables are properties of a geometry object, they are all independent of each other (if one variable is removed there is no impact on the values of the other variables
    - This means the model will still work even if we have very few variable columns

Disadvantages:

- This model is based on the Tobler Python Package (part of the PySAL Geospatial Analysis library) released in 2019 that is still in active development
    - As of the writing of this document, there are only 3 spatial interpolation models available
- This model has no hyperparameters, no coefficients of determination, or any other accuracy metrics. The only way to improve the model is with better data
- The model suffers from the Blackbox Problem, in that it is difficult to analyze the specifics of how it is interacting with the data

- Suffers from the Modifiable Areal Unit Problem (MAUP) (discussed at length in the previous section
- There is an extensive amount of prior ETL that must be done before the data is in the proper format to be used by the model

Improvements to the Machine Learning Model/Tuning

As mentioned previously, there are no hyperparameters to tune and no accuracy metrics to measure the performance of the model. The model was tuned based on the provided uninsurance data at the census tract level. There are 2010 census tracts in the state of New Jersey; however, approximately 170 did not contain any healthcare information for their occupants. As such, we excluded these census tracts from our interpolation model, and in so doing, improving the results of our prediction. This is the reason why our diagrams contain holes; these correspond to missing census tract data.

Beyond the model itself, one way to improve the accuracy of the results is to use more specific information related to the geographic distribution of uninsured persons within each city and factor that into our prediction. As it stands, our model assumes that the uninsured persons for each city (and by extension each census tract) are evenly distributed across the city. If we knew where uninsured people were clustered, we could use the Tobler Dasymmetric Mapping model to produce a more accurate prediction. This model essentially carves out and removes pieces from our city polygons where uninsured people do not live before making each of its predictions.

Another way to improve the results is to use Tobler's model-based interpolation model. Prior to loading into the model, the variables are put into a separate spatial model formula (such as a regression model) that clusters variables together based on their distribution within a source polygon. The spatial interpolation algorithm will then weigh the variables according to each of these clusters when making its predictions. However, this method only works for one variable at a time. Both of the above methods help to solve (or at least mitigate) the modifiable

areal unit problem (MAUP) by introducing specific geospatial data that reflects the spatial variation in our variables of interest.

Conclusion

The Spatial Area Interpolation method from the Tobler Package is a lightweight, yet surprisingly robust method for interpolating spatial data. Despite the package being in its relative infancy, it is already capable of making powerful predictions. In addition, the preliminary work and analysis that we have done can easily be built upon with the incorporation of more specific data at the city level. Hopefully, this report has been an informative introduction to the world of geospatial data analysis, and its many applications in the world of healthcare!