Characteristics of Health Insurance and Access to Care across the United States and Minnesota

Elsa Carlson, Kylee LaPierre, Tarick Mehanna, Logan Sell

Introduction

During the coronavirus pandemic and even some previous years, we have seen some concerning decreases in healthcare coverage. This project report reflects on data gathered from several sources including the U.S. Census Bureau and the Behavioral Risk Factor Surveillance System (BRFSS), from which we have used to analyze healthcare more precisely. Specifically, we examine factors behind lacking health insurance or lacking sufficient access to healthcare. Additionally, we devise a machine learning model which predicts an individual's insurance status based on various demographic factors.

In an increasingly diverse country, it is critical to not only track health insurance coverage overall, but also by various demographic groups. For this reason, a significant portion of our analysis takes into account numerous demographic factors such as age, race, sex, population, income level, etc. Additionally, we attempt to explain the 'why' behind the scenes of health insurance coverage. For instance, we analyze a large collection of data gathered by the National Health Interview Survey, which provides extensive information on not only demographic details but also reasons for lacking health insurance or access thereof. The primary objective of our research is to identify trends or patterns of health insurance coverage so that resources can be best allocated to areas or groups where coverage is less common/accessible. For the purpose of this work, we define *sufficient access to healthcare* as those who do not delay or avoid seeking healthcare services due to cost or location. *Uninsured* refer to those who do not have any type of health insurance.

Exploratory Questions

With these objectives in mind, we formed several questions around this topic describing various aspects we could hone in on using our data sources:

1. How do unisurance rates vary among demographics such as race, age, sex, population, income, and educational attainment level?
2. How does access to healthcare vary at a national level and by county?
3. How does access to healthcare vary among Minnesota counties?
4. Do rates of uninsurance vary based on offered coverage types?

5. Does the rate of uninsurance vary for individuals based on the number of hospitals in their county?
6. Can we predict which areas will have more or less coverage based on demographics, income, hospital count, or other factors?
7. For those who are covered, is their coverage sufficient for their health needs?

Data Sources

All specific data sources used in this project along with a brief description of each source can be seen below. For the citation, see the corresponding source at the end of the document.

1. Business Patterns - Hospitals
    a. This source contains all hospital counts by state in the US for the year of 2019. As well as employee size, establishment type, and establishment count.
2. Demographic and Housing Estimates
    a. This source contains the demographic information for all 50 states plus DC and Puerto Rico. The demographic information is broken up by age, sex, and race. We used the year of 2020 from this source to test our dataset and see how likely someone would be uninsured based on demographics.
3. Characteristics of Health Insurance Coverage
    a. This source contains the demographic information for insured and uninsured (not having health insurance) for the year of 2019 and for all 50 states plus DC and Puerto Rico. The demographics include age, sex, race, education status, annual salary, educational attainment, and if you were native (born in the US) or foreign born (born out of the US). We also used this source for our data for our Machine Learning model for the years of 2015 to 2019 to test our model and 2020 to train our model.
4. Private Health Insurance Coverage
    a. This source contains the private health insurance information for some counties within Minnesota in the year 2019. The private health insurance includes employer-based, direct-purchase, and military along with poverty level, full-time worker, and if they have private health insurance alone.
5. Public Health Insurance Coverage
    a. This source contains the public health insurance information for some counties within Minnesota in the year 2019. The public health insurance includes medicare, medicaid, and veterans assistance along with poverty level, full-time worker, and if they have public health insurance alone.
6. Small Area Health Insurance Estimates (SAHIE) 2019

    a. This source breaks down the insured and uninsured populations by demographic factors such as age, race, sex, and income. These are also broken up by county as well as state totals, allowing us to geographically analyze insurance vs. uninsurance.

7. NHIS
    a. This source's data comes from a plethora of survey questions, containing very thorough and specific insurance details including a variety of demographic details of the insured / uninsured, as well as reasons for not being insured.
8. BRFSS Metropolitan Area
    a. This source contains information for the major metropolitan areas in Minnesota for the year of 2019. It is broken down into different questions and responses as well as topics that relate to health coverage, which is what we looked at.
9. BRFSS Healthcare Access/Coverage
    a. We utilized this source through an API call and it covers all 50 states with questions and categories related to health coverage. We filtered out the data to only the year 2019.
10. States and Counties
    a. This source was utilized for the SAHIE dataset for a list of all counties in the United States.

Data Platform

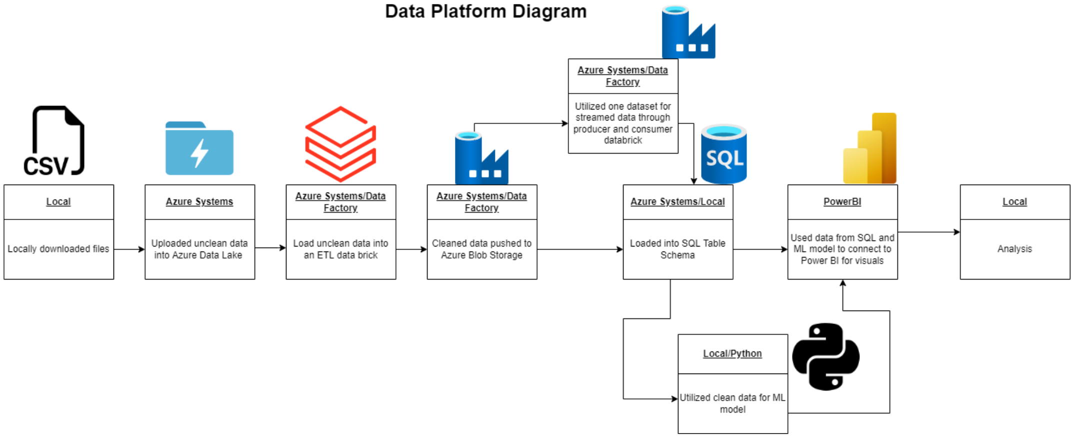
We downloaded all datasets except for the BRFSS National Healthcare Coverage datasets as CSVs in order to clean, normalize, and upload them into a SQL database; we accessed the latter dataset using an API call. We performed the ETL steps in two different databricks, one for general data cleaning & one for loading into our SQL database; we have explained the precise steps used to clean & normalize the data in our ETL Report.

All of our data was preexisting, but we simulated a datastream by running a dataset through a producer & consumer (the dataset in question is S2701: Selected Characteristics of Health Insurance Coverage in the United States). We successfully automated this using a data factory set to refresh every 15 minutes. Each time the factory refreshed, the producer pulled the dataset from an Azure blob and sent the data to the consumer. The databrick containing the consumer received the messages & reorganized the data into an appropriate format, then overwrote the existing table in the SQL database.

To perform our analysis and create visualizations, we created a data flow in PowerBI that connected to our SQL database, once again on a regular refresh schedule. We used this dataflow as our data source in the desktop app, enabling us to perform our analyses locally in reports.

We also created a machine learning algorithm using the non-streamed data, which we will explain later in this report.

All steps for the creation of the dataframe, in order, can be viewed in the diagram below:


Data Platform Diagram
CSV
Local
Locally downloaded files
Azure Systems
Uploaded unclean data into Azure Data Lake
Azure Systems/Data Factory
Load unclean data into an ETL data brick
Azure Systems/Data Factory
Cleaned data pushed to Azure Blob Storage
Azure Systems/Data Factory
Utilized one dataset for streamed data through producer and consumer databrick
SQL
Azure Systems/Local
Loaded into SQL Table Schema
PowerBI
Used data from SQL and ML model to connect to Power BI for visuals
Local
Analysis
Local/Python
Utilized clean data for ML model

After creating visualizations in PowerBI reports, we combined all the visualizations in a dashboard on the PowerBI web application.

Machine Learning

We performed ANOVA tests on the columns in our training and testing data before we did any machine learning to make sure the columns we were using were statistically significant. Our dependent or categorical variable, insurance category, was tested against the rest of our independent or numerical columns. Each numerical column contained the populations of different categories like age ranges, race/ethnicity, gender, place of birth (US or abroad), educational attainment, and salary for all counties and states in the US. We used one categorical column for the ANOVA tests; insurance category. Insurance category was broken into three different criteria; total, insured, and uninsured and we focused on the insured and uninsured categories. Each row of the dataset was broken down into the state and county where people lived as well as their insurance status. We performed the one-way ANOVA tests on the insurance category for the years 2015 to 2020.

From the one-way ANOVA tests, we found that the insurance category had a statistically significant effect on each of the numerical columns. Each p-value was lower than the significance level we chose at 0.05. Thus from the test we concluded that age, race/ethnicity, gender, place of birth, educational attainment, and salary are significant in predicting whether someone was insured or uninsured.

Once we determined that each column was statistically significant in determining if someone was insured or uninsured, we started developing a machine learning model. The model we decided to go with was linear regression as we had a categorical column for insurance category, our dependent variable, and numerical columns for the rest of the data. We started first by creating a dummy variable for the insurance category and getting the data in 0 for insured and 1 for uninsured. We then trained the linear regression model on the years of 2015 to 2019 and tested on the data from 2020. We started with the default hyperparameters for linear regression and hypertuned the parameters to use the solver as liblinear, the penalty as l1, and the C-value as 1.201 as determined by the GridSearch that we performed.

The purpose of the machine learning model is to predict whether someone is insured or uninsured based on their location, age, race/ethnicity, gender, place of birth, educational attainment, and salary. The model ended up being 84% accurate with most people being categorized correctly, about 500 people being categorized as uninsured when they had insurance, and less than 10 people being categorized as insured when they had no insurance.

Analysis

Introduction

We used PowerBI to visualize our datasets. Each group member began by creating multiple visualizations based on the datasets we had processed in previous steps of our project. We utilized PowerBI's desktop application to visually explore these datasets, experiment with formatting options and create the first versions of our visualizations.

After finalizing the general layout and contents of our visualizations, we published them to the PowerBI web service. We decided to organize our group dashboard into five sections: a brief introduction, a section containing visualizations related to health insurance, coverage and access on a national level for the United States in 2019, a section containing visualizations and information related to our machine learning model, a section containing visualizations related to health insurance, coverage and access within Minnesota in 2019, and a final section containing our data sources.

In our introduction we defined access to healthcare and uninsured status, two key terms we used to frame our research. We also included the exploratory questions that have shaped our research since the beginning of our project. Finally, we included information on the timeframe of our data sources. We primarily used data from 2019, and the data we used to develop our model covered the period of 2015 to 2019.

National

We used the national section as a tool to frame the demographic trends of healthcare access and insurance status in the United States as of 2019. We placed this section after the introduction to provide viewers of our work with context for these trends. Because our sources lack data on healthcare access and insurance in many rural and lower-population counties across the country, we decided that a national overview would provide an introductory context that is more inclusive and less skewed toward metropolitan areas than a county-by-county analysis.

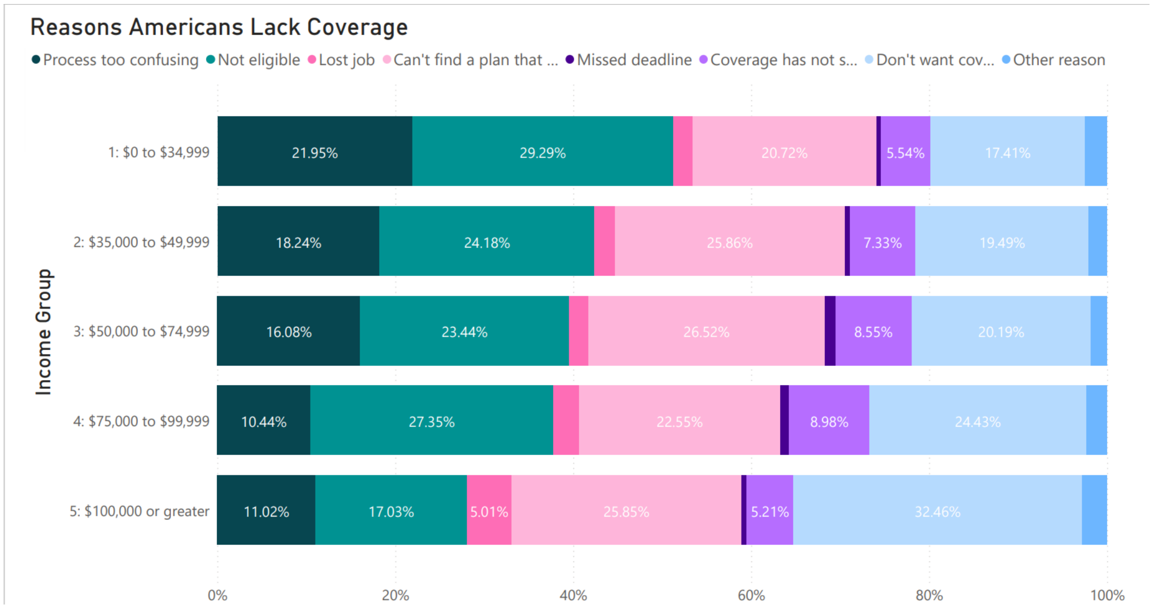
Topics we analyzed in the national section include the influence of cost on insurance status and usage of healthcare resources, breakdowns of insured and uninsured status among different demographic categories, and hospital statistics by state.

We first looked at the self-reported reasons why uninsured Americans lack health insurance. Here are some key takeaways:

- Uninsured Americans of all income levels reported excessive cost as the most common reason for their lack of insurance; in each group this reason accounted for 40% of responses.
- The second-most commonly cited reason for lacking health insurance was not being eligible. This was cited most often by lower- and mid-income Americans.
- Another commonly cited reason by low-income Americans is that the process for signing up is too confusing. This was a common reason across the board, but the number of people citing it steadily decreased as income increased. Notably, there was a noticeable decrease in the number of people who found the process confusing from the lowest income group to the highest (11.18% to 6.40%, respectively)
- "Don't want coverage" was another common reason for lacking health insurance. 17.55% of respondents in the highest income group cited this reason, relative to 11.41% in the lowest income group.

In this survey, respondents could give multiple reasons. The visual below shows the breakdown of common reasons, limited to people who *also* reported that insurance is too expensive with that reason excluded from the visual). This shows that of the people who find insurance too expensive, a significant number of them find the process confusing and can't find a plan that meets their needs.

One step to ameliorate this issue could be to devote more resources to providing educational materials on insurance towards uninsured demographics.



We also looked at two questions on the BRFSS Healthcare survey data: first, whether the survey respondent avoided healthcare in the past year due to cost, and second, whether they have any kind of health insurance. We compared these results to demographics, examining age group, educational attainment, gender, income, and race.

We found similar results for both questions; the group least likely to avoid medical care due to cost was consistently the most likely to have insurance. For age, it was Americans older than 65; for educational attainment, it was Americans with a high school diploma or GED, but no higher attainment; for income, it was Americans with a household income greater than $50,000. While the vast majority of respondents did have insurance, the least likely groups to have insurance include college graduates, 45-64 year olds, and Americans with an annual income of $25,000-$34,999. The racial/ethnic group least likely to have insurance was people who identify as multiracial and not Hispanic, whose rate of uninsurance of 27% was nearly 15% higher than the racial demographic with the second-highest insurance rate. The category least likely to be uninsured were non-Hispanic White Americans.

Unfortunately, the BRFSS dataset, which was aggregated by category, does not show the intersections of any of these demographics. We would like to see more research available regarding educational attainment, income, and race to see how responses to these questions change based on multiple factors. Additionally, there was data missing in some of these categories; for instance, the sample size for Native Hawaiian or Pacific Islanders was extremely low and even absent altogether from some states' data, as was the data for women in some states.

While we could not compare these questions directly using the BRFSS data, we found that the NHIS data was consistent with our analysis of the former. In 2019, uninsured people in the United States reported much higher rates of worry over healthcare cost than those with insurance coverage. Additionally, the NHIS data shows that foreign-born people in the United States were much more likely to be uninsured than people born in the country.

In the Selected Characteristics of Health Insurance Coverage dataset, we found that the highest populations to not have insurance include individuals aged 25 to 44 years old and 65 and older with most insured people belonging to the age group 6 to 18 years old. After looking at age, we were interested in how salary played a part in someone having insurance or not. The highest population of insured individuals had salaries over $100,000 and the highest population of uninsured individuals had salaries less than $25,000. For race, there was some variety between the different categories for the uninsured populations but white or caucasian were the highest population to be insured.

After looking at the different ages, salaries, and races, we then looked at place of birth and educational attainment to see how these might play a role in determining if someone is insured or uninsured within the Selected Characteristics of Health Insurance Coverage. Since there is a high number of the population that is born in the United States, we took a look at the citizens that were foreign born, naturalized, or not a citizen. Of these categories, someone who was foreign born is more likely to be uninsured than someone who is naturalized or not a citizen. Finally, when looking at educational attainment, people with bachelor's degrees or higher are more likely to be insured and people who have a high school degree or equivalent are more likely to be uninsured.

Next, we examined other demographic factors such as location, age, sex, and income, directly against uninsurance rates from the Small Area Health Insurance Estimates dataset. By isolating both age and income level each by sex, we found that males consistently have a higher uninsurance rate than females by anywhere from approximately 1-3%. Looking at trends by age, it appears that this disparity by sex is more prominent in ages 18-40, compared to 40-64. This could be due to several different reasons; more research would need to be done to pinpoint the exact cause.

With regards to income level, we came across an interesting observation; as income level increases, the bigger the gap in uninsurance rate between males and females. To elaborate, there is virtually no difference in uninsurance rate by sex at or below 138% of the federal poverty level but as income level increases all the way up to 400% at the federal poverty level or below, the uninsurance rate will gradually increase for males more than that for females.

Finally, we used the SAHIE dataset to examine uninsurance rates by location. On average, it appears that the Midwest and Northeast have lower percent uninsured populations relative to the West and South, with most states between 5-10%, while the other regions often see states with a 10-20% uninsurance rate.

We then looked at hospital counts and types across the country in Business Patterns - Hospitals dataset. These vary greatly, with most of the hospitals being nonprofit and government owned hospitals. Most of the nonprofits only have 10 to 19 employees compared to the government owned hospitals having 100 to 249 employees.

Minnesota

As an initial overview, we first created a map of Minnesota by the counties available from the census data, showing average percent uninsured by county. While there does not seem to be any significant differences, counties in central Minnesota appear to have lower percent uninsured rates compared to those for northern and southern counties.

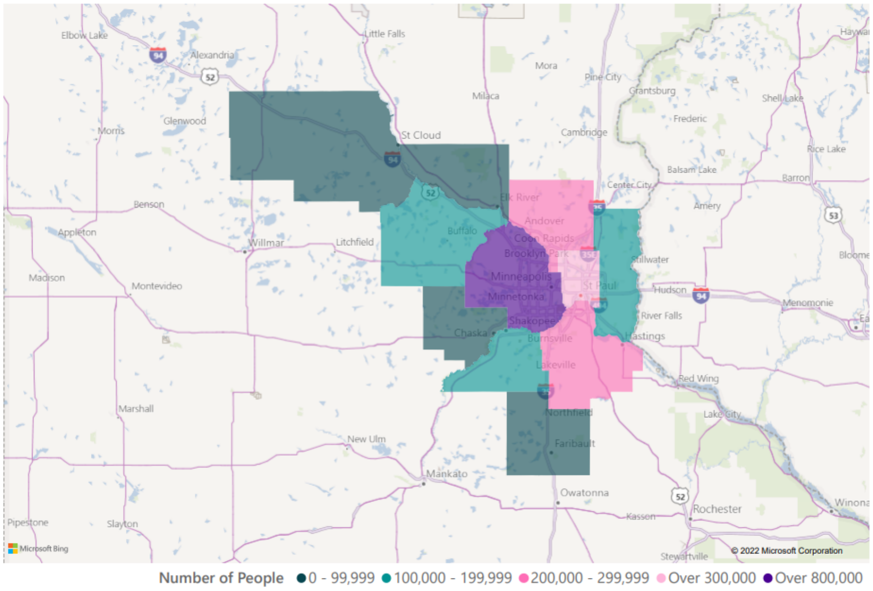
Next, we used BRFSS data on metropolitan statistical areas (MSAs) in Minnesota to visualize information on healthcare access and insurance status in the state's most populous communities. The MSAs included in the survey for 2019 were Minneapolis-St.Paul-Bloomington MN-WI, Duluth MN-WI, St. Cloud MN, Rochester MN, and Fargo ND-MN. BRFSS respondents in each MSA reported low rates of uninsured status: the highest was 10.48% in St. Cloud; in all other MSAs less than 10% of respondents reported a lack of insurance. Similarly, most respondents did not avoid care due to cost over the past 12 months, although St. Cloud lagged behind the other MSAs in this measure as well–11.95% of respondents from St. Cloud reported avoiding care due to cost, compared to 10.21% in Duluth and 10% or fewer in the other three MSAs.

We also used two Census datasets (S2703: Private Health Insurance Coverage by Type and Selected Characteristics and S2704: Public Health Insurance Coverage by Type and Selected Characteristics) to examine health insurance data on a county level in Minnesota. Notably, this data was only available for a select number of counties, namely Anoka, Blue Earth, Carver, Crow Wing, Dakota, Hennepin, Olmsted, Ramsey, Rice, Scott, Sherburne, St. Louis, Stearns, Washington, and Wright Counties, most of which are located within one or more of the MSAs mentioned above.

Overall, we found private insurance (consisting of employer-based coverage, direct-purchase coverage, and military coverage) enrollment to be significantly higher than public insurance (consisting of Medicare coverage, Medicaid and other means-based coverage, and VA coverage) enrollment. 2,621,404 people in the select counties possessed employer-based coverage; the combined enrollment number for the other five types of insurance was 1,957,116 people.

To further examine county-level differences in insurance status, we made four choropleth maps of 11 contiguous counties in the Minneapolis-St. Paul-Bloomington Combined Statistical Area (CSA). The counties included in these maps are Anoka, Carver, Dakota, Hennepin, Ramsey, Rice, Scott, Sherburne, Stearns, Washington and Wright Counties.[1] Two of the four maps examine private insurance enrollment alone and public insurance enrollment alone, respectively. We found the total number of public insurance enrollees, at 593,972, to be slightly greater than one-fourth the total number of private insurance enrollees, at 2,266,433. Hennepin, Ramsey, Dakota, Anoka and Washington counties had the highest number of enrollees in both categories.

Employer-Based Insurance Alone or in Combination in Select Counties, Minneapolis-St. Paul-Bloomington Combined Statistical Area, 2019



Employer-Based Insurance Enrollment

County	Number of People
Hennepin County, MN	803114
Ramsey County, MN	317378
Dakota County, MN	292971
Anoka County, MN	234417
Washington County, MN	178932
Scott County, MN	108173
Wright County, MN	101207
Stearns County, MN	97914
Carver County, MN	76914
Sherburne County, MN	65254
Rice County, MN	39775
Total	**2316049**

Note: Stearns County is also considered part of the St. Cloud Metropolitan Statistical Area.

The Minneapolis-St. Paul-Bloomington CSA counties contain a large proportion of the select county enrollees mentioned previously: 2,316,049 (88.35%) of the 2,621,404 overall employer-based insurance enrollees resided in one of these 11 counties. 1,606,867 (61.3%) of these enrollees lived in Hennepin, Ramsey, Dakota, and Anoka Counties alone.

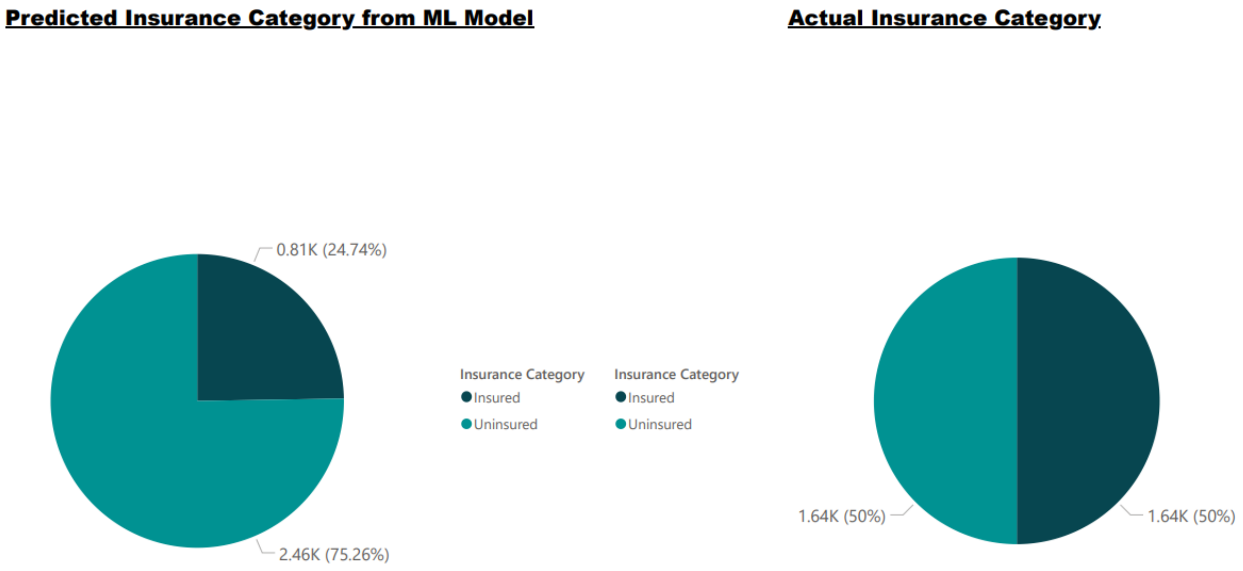
Our fourth map for the CSA examines public insurance enrollees who worked full-time in the CSA counties. We found that 84,936 enrollees fell into this category, equivalent to 14.3% of the

[1] Other counties considered part of the Minneapolis-St.Paul-Bloomington CSA were not represented in our datasets. Stearns County is also considered part of the St. Cloud MSA.

CSA's 593,972 overall public insurance enrollees. 47,518 (55.94%) of these workers lived in Hennepin and Ramsey Counties.

Machine Learning

We used the Selected Characteristics of Health Insurance Coverage dataset for the years of 2015 to 2019 for the training data and the year of 2020 for the testing data. We were interested to see how the insurance categories were determined if the COVID pandemic didn't happen. The machine learning model was 84% accurate at determining if someone had insurance or didn't have insurance based on their location, age, race/ethnicity, gender, place of birth, educational attainment, and salary. From the model, it predicted that 75% of Americans were uninsured in 2020 whereas the actual number was 50%. We would rather have people that are insured be mislabeled as not having insurance than someone who doesn't have insurance be labeled as having insurance since we can give access to resources to people who need it rather than have someone who needs insurance miss the opportunity for additional resources.



Predicted Insurance Category from ML Model

Actual Insurance Category

Conclusion

Through the utilized data sources and explanatory questions to guide this analysis, we narrowed down the broad scope of healthcare to a clean and organized database in order to obtain various insights about how and why people lack healthcare of access thereof; these include insurance vs uninsurance rates, the impact of several different demographic variables, reasons for lacking health insurance, and finally the interpretation of these results.

Our analysis begins by examining healthcare at a national level, where we find numerous potential barriers to health insurance including affordability, confusing sign-up processes, educational attainment, and many more factors. Additionally, we discovered a disparity that favors females over males when it comes to having health insurance, especially in younger age groups and higher income levels.

We then proceed to zoom into Minnesota alone, where we find that counties in central Minnesota appear to have lower uninsurance rates compared to those more northern and southern. Additionally, we discovered that private health insurance enrollment trails significantly ahead of public health insurance enrollment. Finally, we created a model using logistic regression, predicting whether an individual is insured or not based on demographic factors listed previously, with an 84% accuracy rate.

It is critical to track healthcare in specific ways as shown above, whether at a state or nation wide level. Further, machine learning can offer valuable predictions to help understand aspects of healthcare, whether we predict an individual's insurance status or forecast future insurance rates or another aspect. After all, understanding this complex topic is the first step to improving the healthcare system.

Resources

1. Bureau, U. S. C. (2019). *CB1900CBP: All Sectors: County Business Patterns by Legal Form of Organization and Employment Size Class for U.S., States, and Selected Geographies: 2019*. Explore census data. Retrieved September 21, 2022, from https://data.census.gov/cedsci/table?q=hospital+cb&g=0100000US%240400000_0400000US27&tid=CBP2019.CB1900CBP

2. Bureau, U. S. C. (2019). *DP05: ACS Demographic and Housing Estimates*. Explore census data. Retrieved September 21, 2022, from https://data.census.gov/cedsci/table?q=DP05&g=0100000US%240400000&tid=ACSDP1Y2019.DP05

3. Bureau, U. S. C. (2019). *S2701: Selected Characteristics of Health Insurance Coverage in the United States*. Explore census data. Retrieved September 21, 2022, from https://data.census.gov/cedsci/table?q=health+insurance&g=0100000US%240500000_0400000US27&tid=ACSST1Y2019.S2701&moe=false

4. Bureau, U. S. C. (2019). *S2703: Private Health Insurance Coverage by Type and Selected Characteristics*. Explore census data. Retrieved September 21, 2022, from https://data.census.gov/cedsci/table?q=health+insurance&g=0400000US27%2C27%240500000&tid=ACSST1Y2019.S2703

5. Bureau, U. S. C. (2019). *S2704: Public Health Insurance Coverage by Type and Selected Characteristics*. Explore census data. Retrieved September 21, 2022, from https://data.census.gov/cedsci/table?q=health+insurance&g=0400000US27%2C27%240500000&tid=ACSST1Y2019.S2704

6. Bureau, U. S. C. (2022, August 10). *2008 - 2020 small area health insurance estimates (SAHIE) using the American Community Survey (ACS)*. Census.gov. Retrieved September 21, 2022, from https://www.census.gov/data/datasets/time-series/demo/sahie/estimates-acs.html *Only used the year 2019 from this website.*

7. CDC. (2021, April 5). *NHIS - 2019 NHIS*. Centers for Disease Control and Prevention. Retrieved September 21, 2022, from https://www.cdc.gov/nchs/nhis/2019nhis.htm

8. CDC. (2021, December 16). *Behavioral risk factors: Selected metropolitan area risk trends (SMART) MMSA prevalence data (2011 to present)*. Centers for Disease Control and Prevention. Retrieved September 21, 2022, from https://chronicdata.cdc.gov/Behavioral-Risk-Factors/Behavioral-Risk-Factors-Selected-Metropolitan-Area/j32a-sa6u?category=Behavioral-Risk-Factors&view_name=Behavioral-Risk-Factors-Selected-Metropolitan-Area

9. CDC. (2021, September 14). *BRFSS: Table of health care access/coverage*. Centers for Disease Control and Prevention. Retrieved September 21, 2022, from

https://chronicdata.cdc.gov/Behavioral-Risk-Factors/BRFSS-Table-of-Health-Care-Access-Coverage/f7a2-7inb

10. Prescott, R. (2016, June 22). *USA-cities-and-states/US_CITIES_STATES_COUNTIES.CSV at master · Grammakov/USA-Cities-and-States*. GitHub. Retrieved September 26, 2022, from https://github.com/grammakov/USA-cities-and-states/blob/master/us_cities_states_counties.csv