**Using Social Determinants of Health Variables to Predict Uninsurance Rates in the United States Using Machine Learning**

DS 325 Final Project Report

Spring 2025, Prof Roth

Madison Mariani and Natalie Hall

**Introduction / Abstract:**

Economic, demographic, and geographic variables contribute to the type of insurance coverage an individual has. To better understand certain patterns present in coverage, or lack thereof, we used data from the 2020 Social Determinants of Health (SDOH) dataset, which reported a variety of health-related variables for the US population. Preliminary data exploration consisted of examining how age and state income level (low, middle, and high) impacted insurance type. The exploration was as expected: the higher the income a state had, the higher private insurance rates were, and poorer states relied the most on Medicaid. Middle-income states had the highest rate of uninsurance, which was an interesting pattern in the data. The utilization of Medicare was also as expected when comparing all age groups to individuals under 64, as unexpectedly, as the latter group was not yet eligible for Medicare. Although the rate of insurance type is important, many of the patterns were as expected in the preliminary exploration. Rather than diving deeper into patterns we could already assume to be present, we chose to further explore factors contributing to uninsurance rates. Uninsurance rates are a huge issue in the United States, and typically, vulnerable populations are affected by this most often. Therefore, we hypothesized that uninsurance rates in states across the US could be predicted by age, employment status, mean distance to nearest Emergency Department, income, and education level.

Using the 2020 Social Determinants of Health (SDOH) dataset, we conducted preliminary analyses, tested feature importance using a correlation matrix and a PCA, used a Random Forest to observe predictive power of chosen features, and performed cross-fold validation to assess. We found that the Random Forest model using regional identifiers such as state and region worked best, achieving an MAE of 1.91 and R² of 0.63. Our models and analysis concluded that in addition to age, employment status, mean distance to nearest Emergency Department, incom,e and education level, geography plays an important role in predicting uninsurance rates.

**Methods:**

We used the 2020 Social Determinants of Health (SDOH) dataset, which contains 682 health related variables for every state and respective counties, in the United States. The health related variables within the dataset cover different socioeconomic, demographic, and healthcare access indicators. During preprocessing, we explored the influence of age and income on coverage. As expected, private insurance was associated with state income, where states with average higher incomes had a higher rate of private insurance coverage over other insurance options such as Medicaid, which had a higher enrollment rate in states with lower average incomes. Additionally, Medicare usage was lower for the group under 64 as that group is not eligible and therefore cannot yet enroll in Medicaid.

In addition to age and income, the features initially selected were employment status, education, and mean distance to the nearest Emergency Department. Choosing these five features was built on the basis that insurance enrollment is heavily based on ability to pay, availability through one's employer, ability to utilize resources covered by the plan, among many other factors. Using the features listed previously, we aimed to predict the percentage of uninsured individuals under specific economic and demographic conditions.

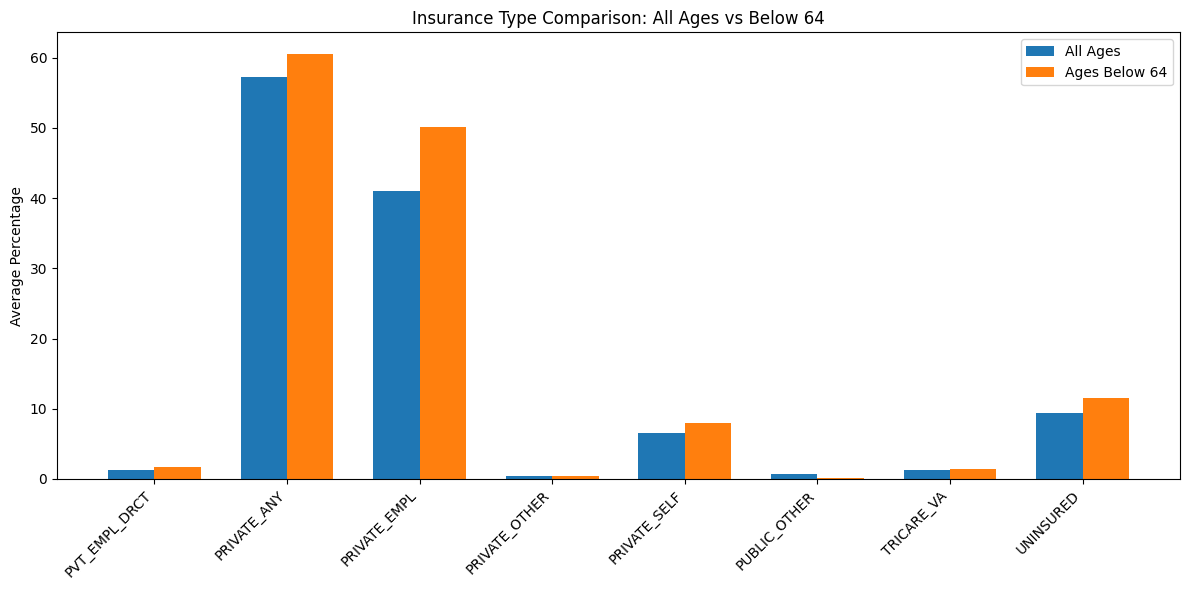
After our preliminary data exploration, we dropped any missing data within the dataset and scaled the data, using StandardScaler. We used PCA to reduce the data's dimensionality and make it easier to visualize. We trained and compared decision tree, gradient boosting, and random forest models, with random forest (RandomForestRegressor) being chosen because it is able to balance accuracy with the interpretation of key indicators. The random forest model initially performed very poorly and to correct, we implemented geographical location to aid in prediction power. We one-hot encoded state and region and incorporated it into the model.

Finally, we validated our model using 3-fold cross-validation to assess generalizability. All analysis was done using Python libraries including pandas, scikit-learn, and matplotlib. GenAI was used to explore reasons behind performance issues and assist in categorizing states into income tiers during our exploratory analysis.

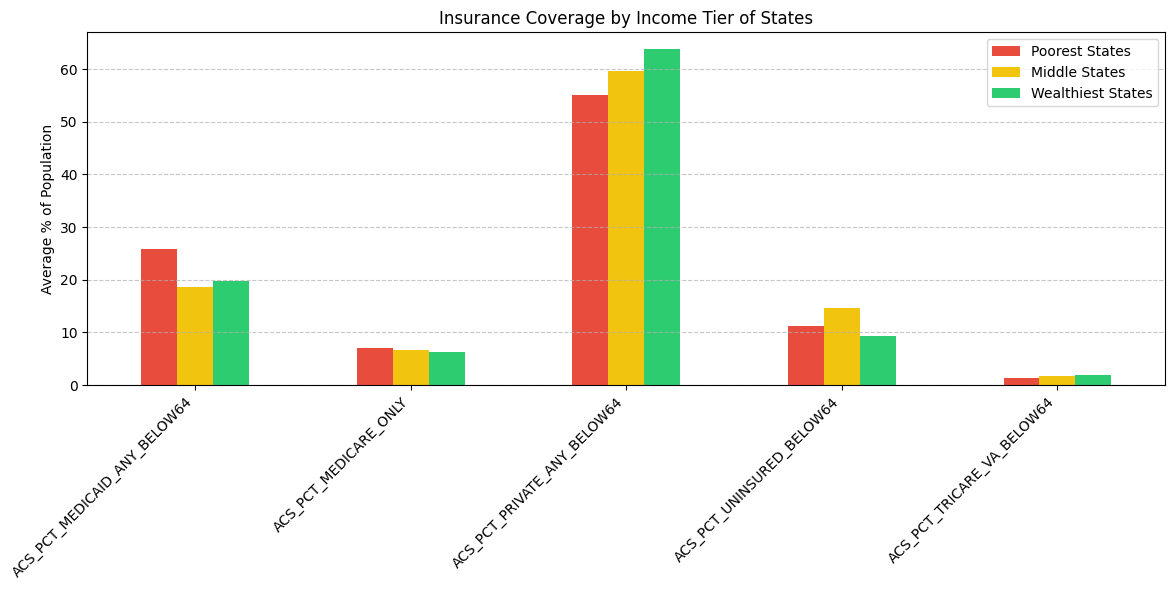
**Results:**

Prior to model training, we examined two key demographic interests in the data: age and income tiers. **Figure 1** shows how insurance type varies between the entire population and individuals under age 64, revealing higher uninsurance rates among the latter due to Medicare ineligibility. **Figure 2** displays insurance coverage by state income tier, where middle-income states had unexpectedly high uninsurance rates. These observations informed our decision to continue analysis with age and income indicators in the model, as well as the additional features (education level, employment status, and mean distance to an ED).

We found that the most important features in determining rates of uninsurance were the percentage of people with less than a high school education, those with income less than $25,000, employment status, and mean distance to the nearest Emergency Department. Several machine learning models were compared to determine the strongest model to predict uninsurance using county-level SDOH variables as recorded in **Table 1**. Our baseline random forest model performed poorly with a R² of only 0.01. However, after incorporating geographical features, the random forest model returned an R² of 0.63. After this, we applied 3-fold cross-validation to the geography-inclusive model to test for generalizability. The performance was weaker under cross-validation with an R² of 0.25.



**Figure 1**: Insurance Type Comparison for All Ages vs Below 64



**Figure 2**: Insurance Coverage by Income Tier of States

| **Table 1:** Comparing Different ML Models in Predicting Uninsurance based on SDOH Data | | |
| --- | --- | --- |
| **Model** | **MAE** | **R²** |
| **Baseline Model** | 3.42 | 0.01 |
| **Geography-aware Random Forest** | 1.91 | 0.63 |
| **Cross-Validation** | 3.04 | 0.25 |
| **Gradient Booster Regression** | 3.04 | 0.40 |

**Discussion:**

Our results showed us that our hypothesis was correct by proving that geography plays a major role in influencing rates of uninsurance. Adding state and region into the model helped us account for systematic things such as policy, infrastructure, and cultural differences across the United States. We found that our best model was Random Forest with geography-aware features. This performed a lot better than some of the simpler models because it reduced average error and explained over 60% of the variance in the model.

Level of education, household income less than $25,000 and being unemployed turned out to be the biggest factors affecting uninsurance rates. Most surprisingly, we found that middle-income states had the highest levels of uninsurance. We concluded that this was due to an issue of the residents of these states earning too much to be on Medicaid but too little to be willing to spend money on health insurance.

Some challenges we faced when doing this project included our geographic data being inconsistently formatted and having high dimensionality which resulted from one-hot encoding. Gradient boosting worked well but not as well Random Forest in some aspects such as accuracy and stability. Lastly, we used PCA to give us an exploratory visualization but not for model input.

This model can be used practically by policymakers to specifically locate counties that have high uninsurance rates and prioritize them when it comes to interventions such as subsidies or outreach programs. Future work can be done by adding on things such as urban-rural classifications or medicaid expansion indicators. This model shows how uninsurance rates are affected by things such as structural economic and social factors, not just an individual's own actions.

**References:**

"US States in Income Groupings" prompt. *ChatGPT*, 25 Sep. version, OpenAI, 2 May. 2025, [chat.openai.com/chat](http://chat.openai.com/chat).

"How can I adjust this Random Forest to have better predictive power" prompt. *ChatGPT*, 25 Sep. version, OpenAI, 2 May. 2025, [chat.openai.com/chat](http://chat.openai.com/chat).

Agency for Healthcare Research and Quality. (n.d.). *Home*. U.S. Department of Health and Human Services.<https://www.ahrq.gov/>