US Maternal Mortality - Executive Summary

Team: Solange Aveiro, Kristina Callaghan, David Clancy, Stephanie Egler, Fei Li, Katherine Martin.

Motivation: The United Nations' Sustainable Development Goal Target 3.1 is to reduce the global Maternal Mortality Ratio (MMR) to less than 70 deaths per 100,000 live births by 2030. Between 2000 and 2015, while 130 countries significantly reduced their MMR, only five countries had a significant increase in the MMR—with the United States experiencing the greatest increase [1]. Therefore, our goal is to identify key predictors of maternal mortality to inform policies that will decrease the U.S. maternal mortality ratio.

Data Sources:

- CDC Wonder [2]: Contains both data on pregnancy-related deaths, births, and information on the mothers.
- CDC DNPAO [3]: Data from the Division of Nutrition, Physical Activity, and Obesity in the CDC containing behavioral information of the women.
- American Community Survey [4]: Contains data related to income distribution in each state.

The data consists of various economic, health, and pregnancy-related related information in the aggregate for each state from 2010-2019. For privacy reasons, the CDC suppresses data if between 1 and 9 (inclusive) people are affected. The data consists of 19 states (AZ, AK, CA, FL, GA, IL, IN, MD, MI, MO, NJ, NY, NC, OH, OK, PA, SC, TX, WA) and 50 features. The features describe: Income information for families in the state (10), Obesity and Activity information (4), stress-related births (5), method/location of delivery (8), marital status of mother (2), start of prenatal care (12), duration of pregnancy (8). This data was contained in .txt files and .csv files that were parsed, cleaned, and merged into a coherent .csv file for subsequent analysis.

Primary Variables/Features of Interest

The primary features of interest, as shown in the feature importance charts, include:

- **Gestational Age Categories**: Features like Birth 20-27 Weeks and Birth 28-31 Weeks are critical predictors of preterm birth outcomes.
- **Prenatal Care Initiation**: Variables such as Prenatal Unknown Initial Care, Prenatal 4th Month Initial Care, and Prenatal 3rd Month Initial Care highlight the timing of prenatal care initiation.
- **Delivery Methods**: Cesarean Delivery and CNM Delivery represent different delivery methods impacting birth outcomes.
- **Medical Conditions**: Chronic Hypertension Births indicates the presence of medical conditions affecting birth outcomes.
- **Socioeconomic Factors**: Income levels such as income 75,000 to 99,999 and income 50,000 to 74,999 are socioeconomic factors influencing healthcare access and outcomes.

These features can be visualized using the provided feature importance charts, which highlight their relative impact on the model's predictions.

Modeling Approaches

Our modeling approach involved trying different combinations of scalings and regression models. We built a pipeline for various scalings, models, and model parameters and then used scikit-learn's GridSearchCV and a 5-fold cross-validation to determine which combination produced the lowest mean squared error. The regression models included GradientBoostingRegressor, KNeighborsRegressor, RandomForestRegressor, Ridge Regression, SVR, Lasso Regression, Elastic Net Regression, Linear Regression, DecisionTreeRegressor, and a baseline model that computed the average of the target.

Once the best model was identified, we applied it to the testing data and computed its mean squared error. We also applied the baseline model to the testing data as a comparison for our model's testing performance. The best model's top features of importance were also identified.

Accuracy and Strengths/Weaknesses

Of the models trained, the GradientBoostingRegressor with StandardScaler and a learning rate of 0.1, max depth of 3, and 100 estimators performed the best on the training data. On the testing data, the best model achieved a MSE of ~97 and significantly outperformed the baseline model's MSE of 182. The top features of importance in the best model included unknown or no prenatal care, preterm deliveries of 28 - 31 weeks, and a lack of physical activity. Features of little to no importance included receiving prenatal care and a high reported income. The fact that the relative importance of features is consistent with known risk factors for adverse birth outcomes demonstrates the strength of our model.

The greatest weakness of our models was the limited data. The data sources we had suppressed data from regions with smaller populations, so we were often forced to disregard these smaller regions and only consider larger regions in our data collection.

Also, even with this model yielding the lowest MSE over our test set, this is still a very high MSE. This is because we were predicting values (maternal mortality death rates) and it is hard to get a specific value accurately. One thing that could fix this is if we were able to get a classification model to determine high versus low risk categorically. We looked into this and there weren't predetermined categories that would make sense for us to use, so if we were to implement this we would need to determine those categories for ourselves.

Next Steps

Transition from regression-based to classification models by exploring advanced classification techniques; integrate post-pandemic data to comprehensively evaluate COVID-19's impact on maternal health outcomes, focusing on changes in healthcare access and resource allocation; analyze the influence of socioeconomic status, including economic and educational levels, on maternal health, identifying potential barriers and opportunities for intervention; assess patient behavior by examining lifestyle and health behaviors to determine their effects on maternal health outcomes; investigate geographic factors by studying the effect of residential location on access to care and outcomes, considering urban vs. rural disparities; delve into racial disparities to investigate higher maternal mortality rates among Black women, aiming to identify underlying causes and develop actionable solutions; conduct maternal care analysis to examine the relationship between the quality of care during pregnancy and maternal mortality rates, highlighting areas for improvement in healthcare delivery and policy.