Predicting Preterm Birth Using Big Data and Machine Learning Techniques

Negar Darabi

2/7/2020

What is Preterm Birth? Why it is an important issue?

- Preterm birth (PTB): is defined as a birth before 37 weeks of pregnancy.
- It is the leading cause of infant mortality in the U.S.
- In 2013, PTB accounted for 36% of U.S. infant deaths in their first year of life.
- In addition to the monetary cost of PTB, which exceeds 25 billion dollars annually and these babies may suffer from life-long deficiencies.

Objective and Research Questions

I want to investigates the risk factors of preterm birth (PTB) in the largest obstetric population that has ever been studied in this field.

I would like to address two research questions:

- What risk factors are the most important predictors of preterm birth? More specifically, what combination of risk factors increases the chance of preterm birth?
- What is the chance of a preterm birth given the presence of a number of risk factors?

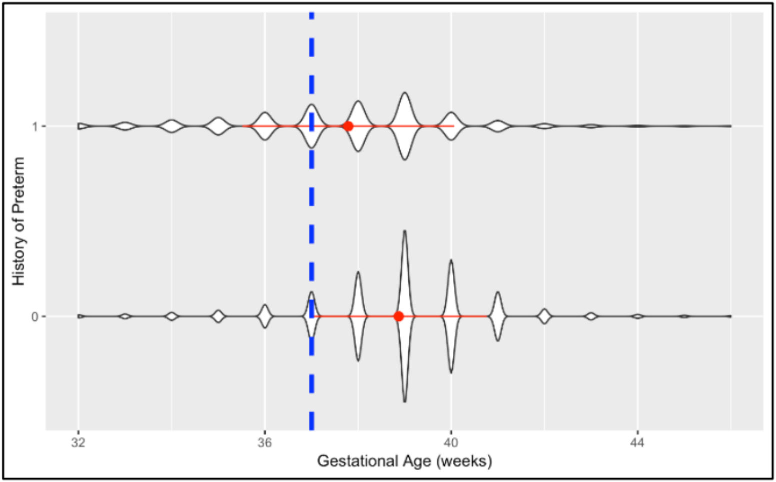
Data Source

- The 2016 U.S. linked birth dataset is obtained and combined with two other area- level datasets, Area Health Resources File and County Health Ranking.
- A cohort of 3.4 million singleton deliveries
- All datasets were linked using a common geographical identifier, the FIPS county codes.
- I performed the data cleaning and preparation in STATA 14.0 and the processing has been coded in R 3.6.0.
- These datasets have a total of 300 variables. After the initial data cleaning, 77 influential variables remained.
- These variables can be classified into the categories of geographical, behavioral, and demographical information of parents, as well as mothers' health status (e.g., history of hypertension, diabetes, sexually transmitted disease, and body mass index).

Visualization

- Data visualization is a challenging task in this study due to the large number of observations.
- We used Violin graphs from the ggplot2 package in R to plot the data and gain more information about the features and their relationship with preterm birth.

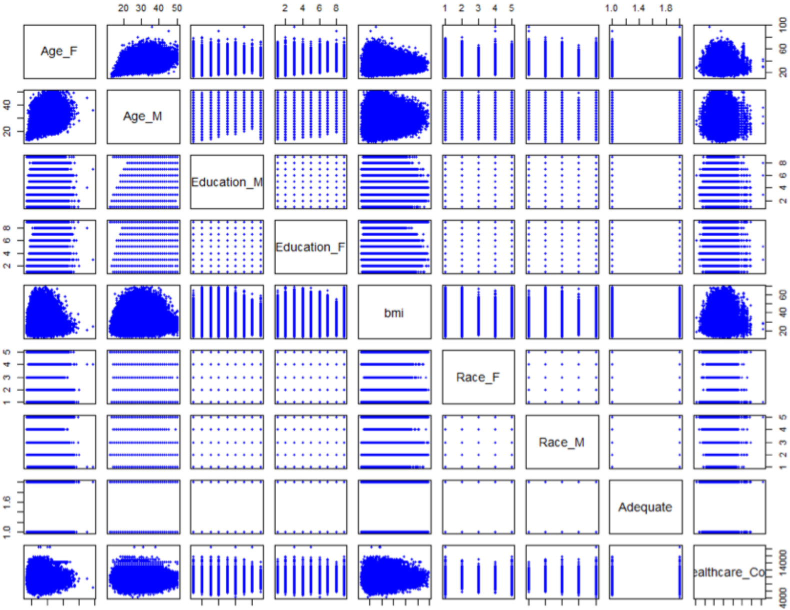
Violin Graph



Method

- The distribution of the response variable is imbalanced. Preterm birth in singleton pregnancies occurs only in eight percent of the deliveries and the remaining are full-term.
- Multicollinearity issue within the dataset.

Multicollinearity



Method

- we are interested in finding the significant interactions among the variables.
- our dataset has 3.6 million records with 77 variables and 20 categorical variables.

Based on these four characteristics, I will apply random forest, gradient boosting machines (GBM), and XGboost

Takeaway (Who gets benefit from the results?)

- Reducing the total PTB rate will results in billion dollars cost saving for the healthcare system.
- It also, prevents life-long deficiencies for preterm babies.
- The result of our work would be helpful for physicians and practitioners with specialty in ObGyn as well as policymakers at both state and national level.