Final Project Draft

Maria Gilbert and Anuka Revi

4/23/2021

Predicting Birthright for Babies

Abstract

summary of our questions, methods, results, and conclusions in a few 100 words

Introduction

Baby announcements are sweet and memorable. Expecting parents usually know their approximate due date at the beginning of the pregnancy and learn about their baby's gender at around 20 weeks mark. While gender reveal, time of birth and name of the new baby are all exciting and great news for the entire family to have, OB-GYNS mostly look at baby's development, growth and their weight. Estimating baby weight in particular is very important, since a large size baby can pose serious injuries to the baby and its mother. Ultrasounds are often used to estimate baby's weight, but they are not always accurate. When I was over 40 weeks pregnant, my doctor told me that the baby was an average size and I didnot need to be induced, however my daughter was born over 8.9lbs which was significantly larger than the estimated size and it made my delivery more complicated. So While exploring my project ideas, I came across to the NBER natality birth data from 2018 that contained the features of the baby like birth date, gender, medical conditions of the birth mother, number of children born etc. My goal is to predict the birthright of the baby depending on the week of pregnancy. Better estimations used by doctors should mean easier deliveries and less complications and its interesting to see how machine learning tools could help us estimate the birthweight.

Methods

My dataset birthweight18.csv contains 20,000 randomly chosen samples, for births that occurred between 37 and 39 weeks of gestation, from 2018 U.S. data files (https://www.nber.org/research/data/vital-statistics-natality-birth-data). There are 39 variables that were chosen from the initial 142 variables on the dataset. These include a unique identifier, birth month (1-12), birth time, sex, birthweight (grams), age of the mother (years), whether or not the mother was born in the U.S., whether or not the baby was born in a hospital; the race, marital status, and education level of the mother; the age, race, and education level of the father, the number of prior live births, the number of prior term births, whether or not the mother was on WIC during pregnancy, indicators for whether the mother had at least one, two, or three cigarettes during pregnancy, the mother's height (inches), the mother's BMI, the mother's pre-pregnancy weight (lb), the amount of weight the mother gained during pregnancy (lb), whether or not the mother had gestational diabetes, hypertension eclampsia, infertility treatment, a previous cesarian section birth, induced labor, anesthesia, ruptured uterus, and/or perineal laceration, whether the birth was vaginal or c-section, whether the birth was attended by a medical doctor or a midwife, and total number of weeks of gestation. We are excluding cases of multiples (twins, triplets, etc.) and pre-term births.

We are hoping to create a model that can predict a baby's weight based on factors that would be known prior to starting labor, and a separate model that can predict a baby's weight based on factors that would be known upon live birth, prior to the actual weighing of the baby. So far, we have worked to clean the data and figure out which variables are the most correlated with birth weight. I was able to get a random forest model working that included weight gain, pre-pregnancy weight, BMI, total gestation weeks, mother's education, father's age, mother's age, prenatal visits, birth month, father's education, father's race, sex, mother's race, mother's marital status, whether labor was induced, previous cesarian, anesthesia, gestational diabetes, mother's birthplace, doctor versus midwife, perineal laceration, infertility treatment, and hospital versus not (in descending order of variable importance), however I could only get the model to work in an R script, it would not knit on R Markdown so we are still working on that.

Data Exploration