Outline

* Project Goals
  + Description
  + Importance of problem
* Exploratory Analysis
* Solutions - each solution should include:
  + Features used and why
  + Classifier used
  + Summary of model results
  + Feature importance analysis

**Project Goals**

In this study, we hoped to understand and predict the likelihood of complications for U.S. births. We define complications at birth to be a delivery that requires a Cesarean section, the presence of an infection, or birth weights that are significantly lower than the mean. Research has suggested that multiple maternal factors are responsible for influencing birth weight of newborns such as age, weight, and previous health history. Researchers have even developed an Artificial Neural Network (ANN) capable of predicting birth weight with 100% accuracy using predictors like these (Al-Shawwa & Abu-Naser, 2019). Similarly, birth weight can be predicted using other measurements from the mother, such as symphysio-fundal height/abdominal girth measurements. These methods are particularly useful in underdeveloped areas where modern healthcare technology is not available. These measurements were found to be highly correlated to the birth weight of newborns (Ademowore et al, 1989). Furthermore, research has suggested that women who experience vaginal bleeding one or more times during the first trimester of pregnancy are more susceptible to underweight newborns. On average, full-term babies born to women with more than one day of bleeding in the first trimester were roughly 3 ounces lighter than the babies of those who experienced no bleeding (NICHD, 2018). These studies have offered motivation for our research towards predicting birth weight.

The data explored comes from the US Births (2018) dataset on Kaggle, which is directly sourced from the Center for Disease Control (CDC) website.

Birth complications is not a straightforward response variable, so we began our exploratory analysis using ‘low birth weight’ as a proxy for complications. We found low birth weight helpful to explore the data, but also found that other common and serious complications were oppositely affected by some of the predictors of low birth weight. So we decided to break apart complications into 3 major response variables: low birth weight, C-section delivery, and onset of infections during pregnancy.

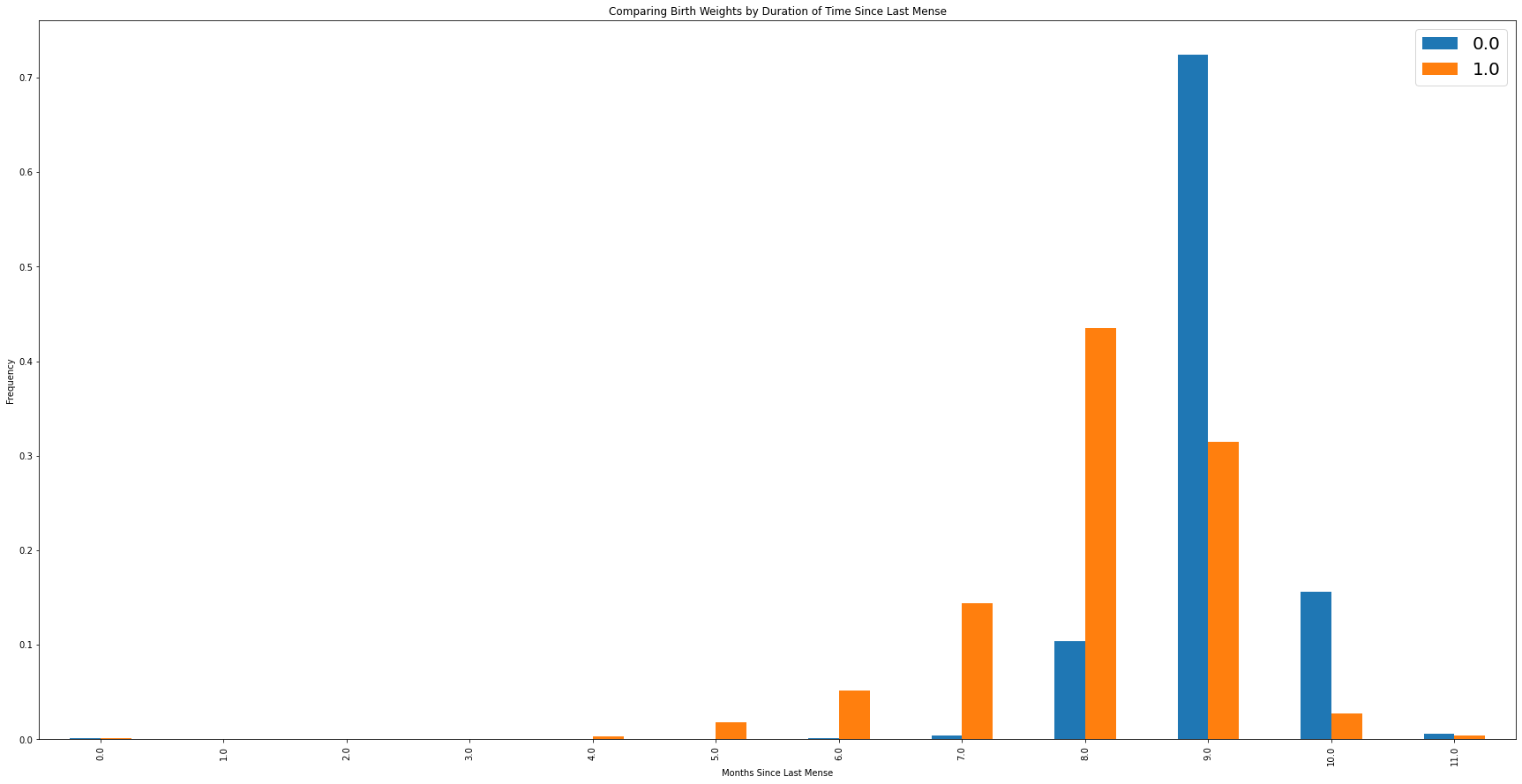
One of the key indicators of good health outcomes during birth is being able to quickly diagnose issues that may occur and then promptly and effectively treat them before they escalate. By creating models that can accurately predict the risk of possible complications of a birth, hospital systems can more accurately forecast patients who are at a high risk of having a complication and resultantly prepare both the patients and providers. Shared decision-making is becoming more prevalent in medicine and with improved information disseminated to patients and healthcare teams symmetrically, patients and doctors can make data-driven decisions to improve outcomes.

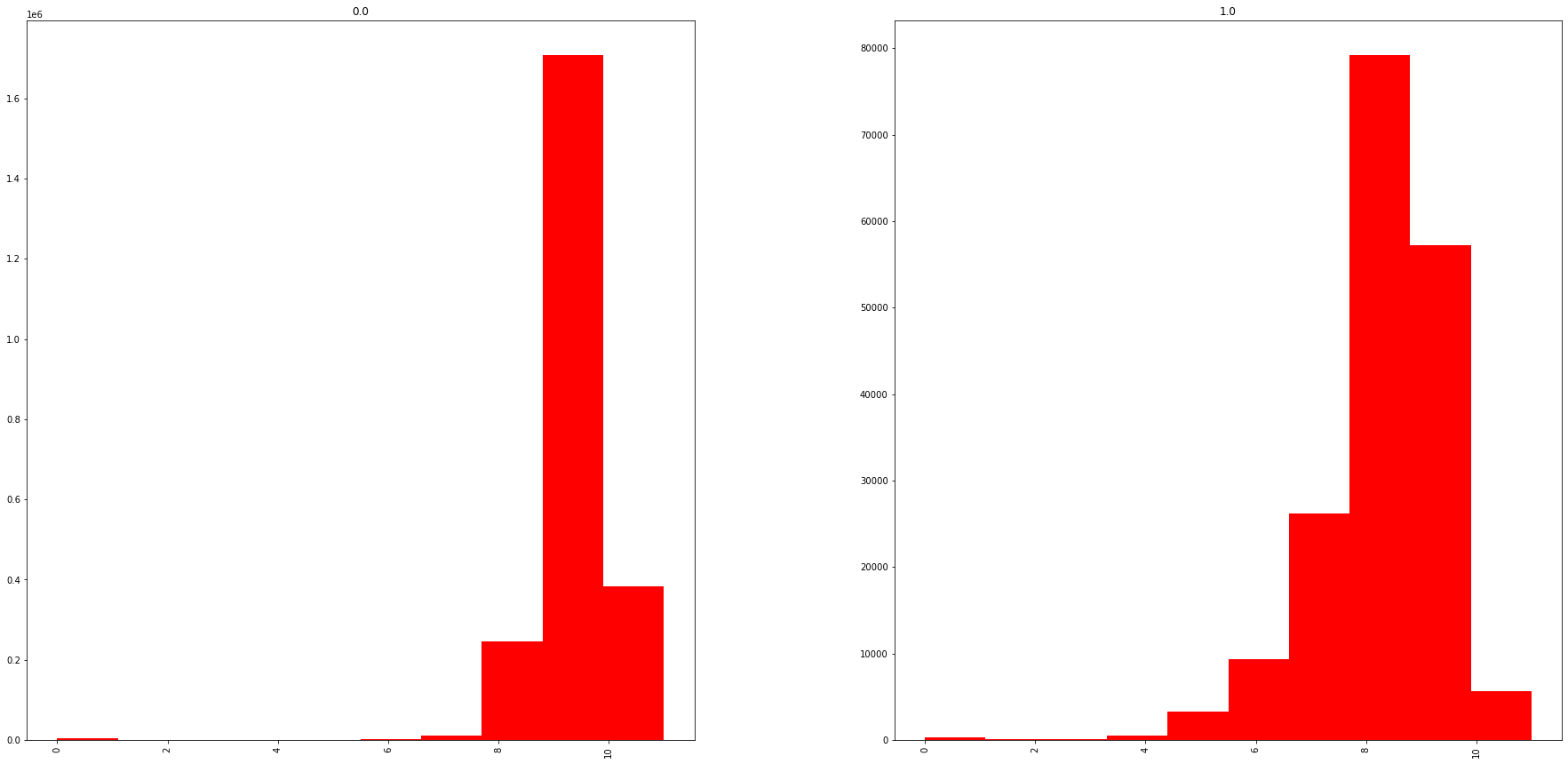
**Exploratory Analysis**

The dataset contains 3,801,534 births and 55 different features that detail different aspects of the mother and newborn’s hospital visit. We first performed exploratory analysis to see how the frequency of normal and underweight newborns are distributed using different predictors.

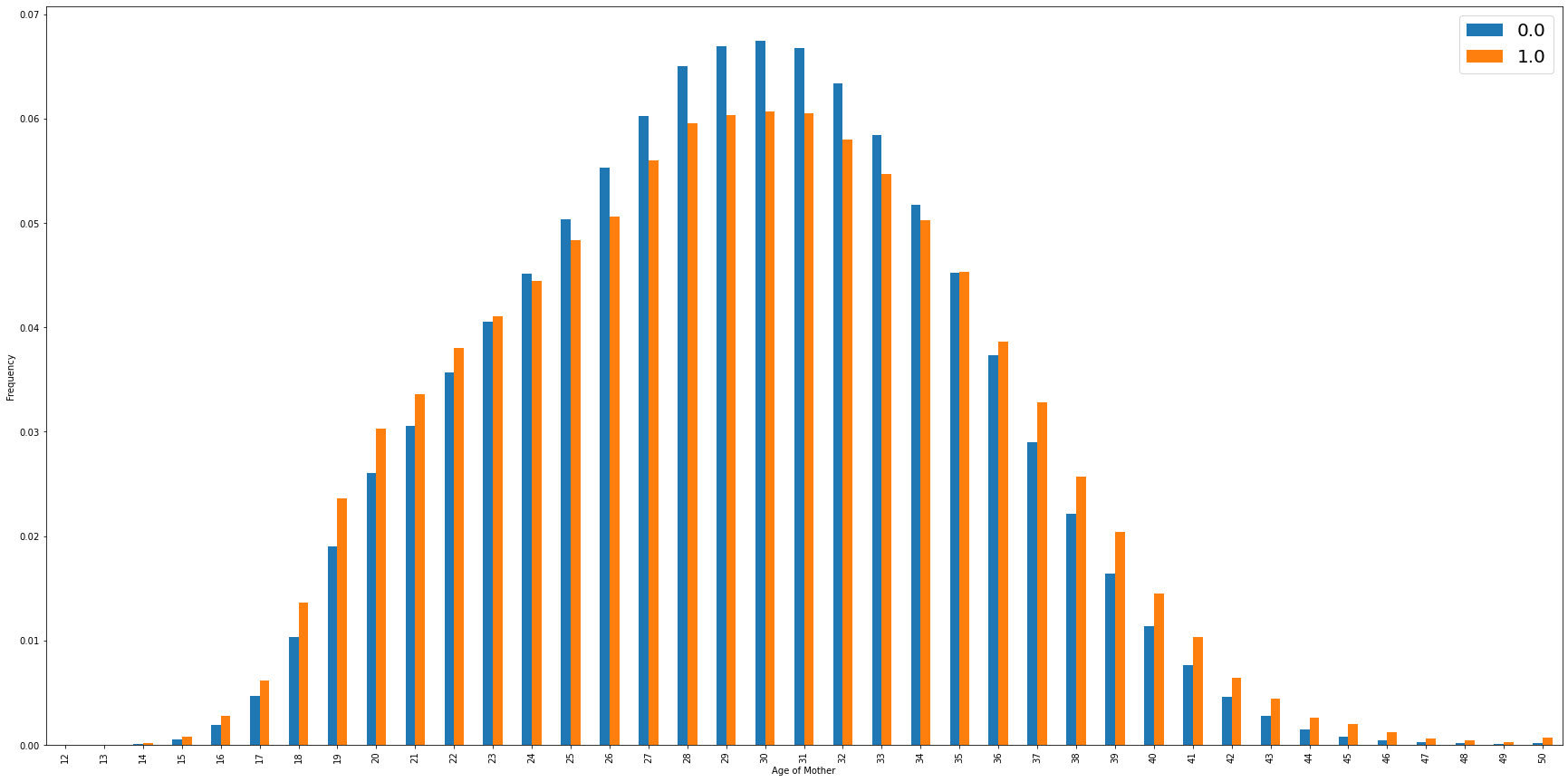
A new column was created to measure the number of months since the mother’s last menstruation cycle and the birth of the child. This was calculated by finding the difference between the baby’s month of birth and the mother’s last recorded menses. We can see that mothers who maintained a regular menstrual cycle past one month into the pregnancy were much more likely to give birth to an underweight baby.

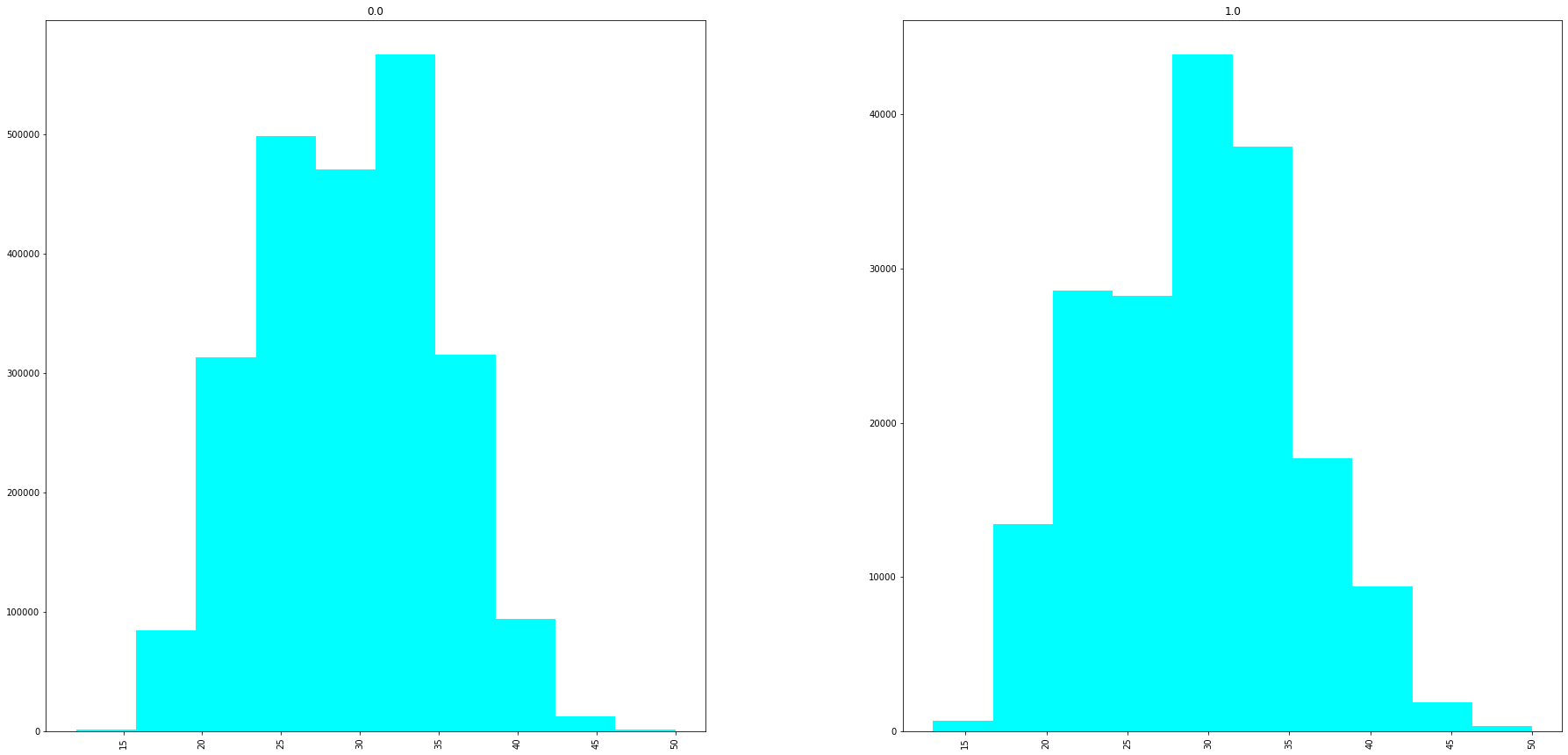
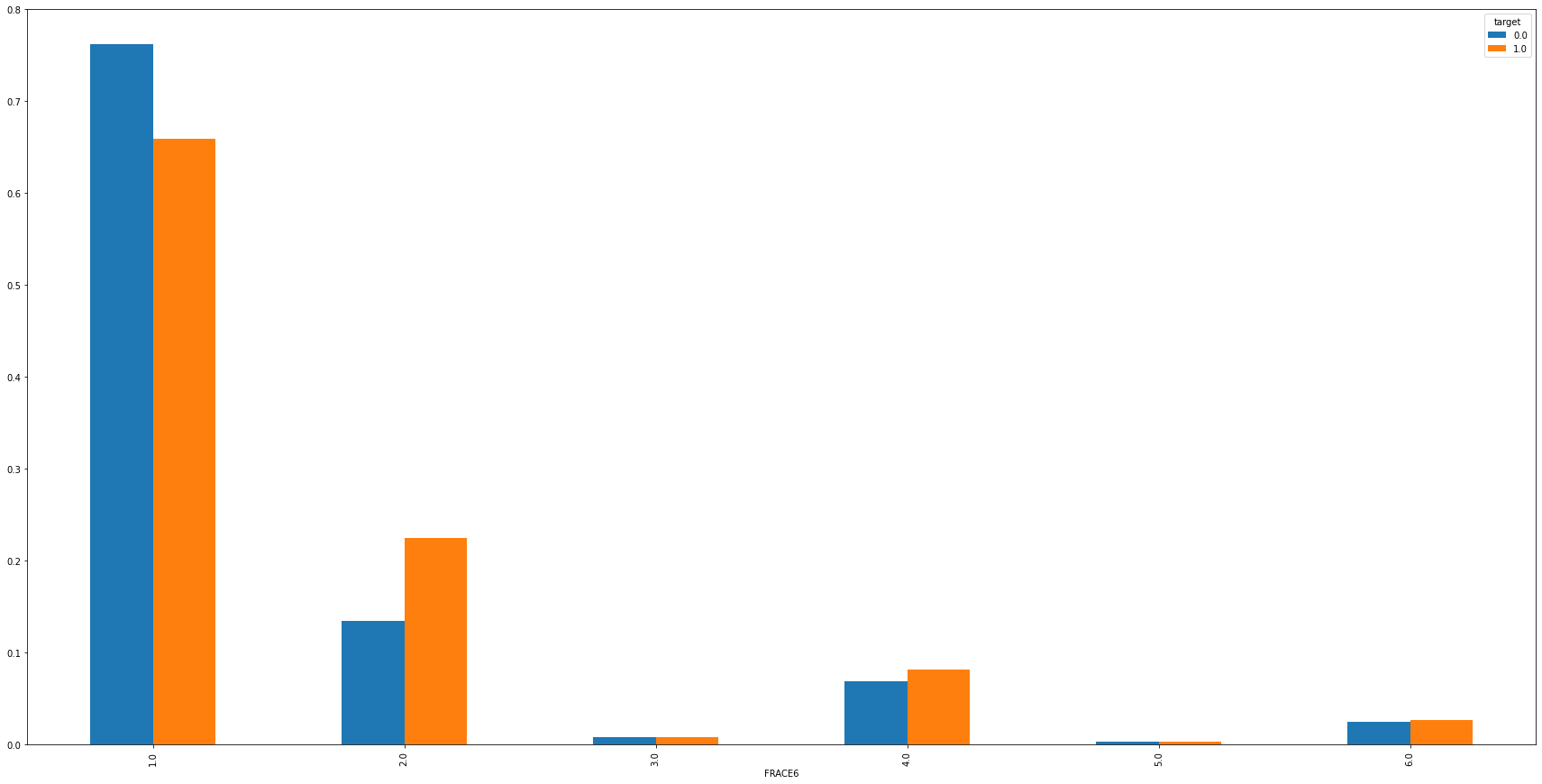
##relate this back to research cited prior



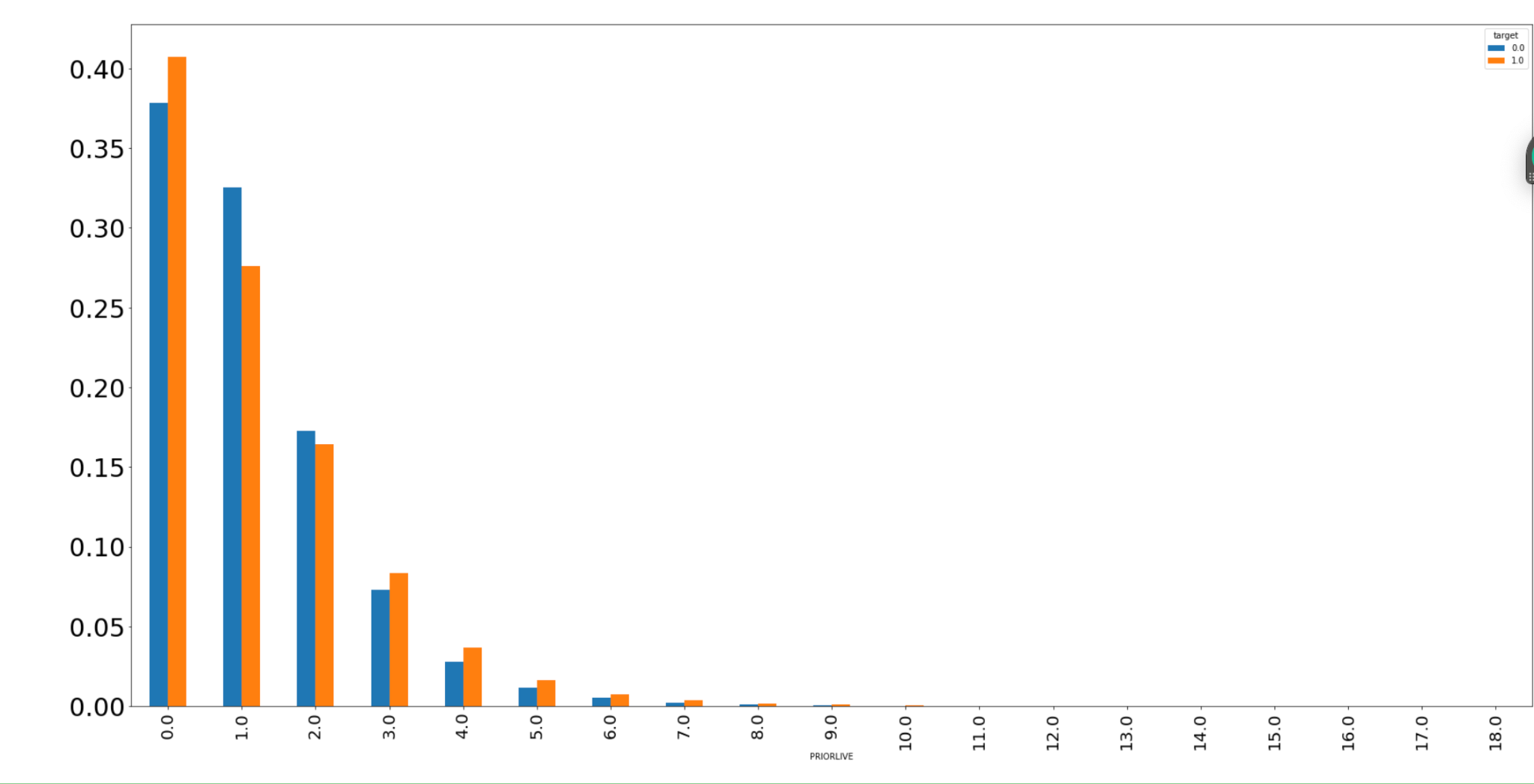


Additionally, mothers below the age of 24 and above the age of 35 had more underweight than normal weight babies.



  
  
When examining the frequency of births for different races of fathers, white fathers are the only group that had more normal weight than underweight newborns.

A new mother was more likely to give birth to an underweight baby, but mothers with 1 or 2 previous births had a higher frequency of normal weight newborns. Mothers of more than 3 previous births experienced higher rates of underweight births than normal births.



**C-Section Analysis**

1 in 3 women in the US deliver via C-Section (up from 1 in 4 over the past 20 years). But the WHO recommends that countries do not have any medical necessity to exceed 10-15% rates of C-sections, putting current US numbers at more than double. From a medical standpoint, healthcare providers might recommend a C-section if:

* Labor isn't progressing normally.
* The baby is in distress.
* The baby or babies are in an unusual position.
* You're carrying more than one baby.
* There's a problem with the placenta.
* There's a health concern for the mother.
* There's a blockage.
* You've had a previous C-section or other surgery on the uterus. It is possible, but rare that women have a natural birth after having a C-section.

There are additional cases where women request to have an elective C-section for their first delivery. Elective C-sections are usually to avoid labor or the possible complications of vaginal birth or simply to plan the time of delivery. However, according to the American College of Obstetricians and Gynecologists, physicians do not recommend this, especially if a woman plans to have multiple children. C-sections can introduce risks to babies and mothers, as well as to future pregnancies.

After investigating the data with a random forest, the most significant variables that impacted the risk of a C-section were the mother’s BMI before pregnancy and the mother’s age during delivery. Women of higher ages and higher BMIs were at more risk of a C-section. When investigating the data with 28 predictors in a logistic regression classification, the most significant predictors were also identified to be BMI and maternal age.

The baseline rate of C-section deliveries is about 32%, meaning that baseline accuracy is 68% (if predicting no C-sections). However, there are no practical benefits from predicting the negative class for all births trying to prepare for a delivery. From a practical perspective, the best baseline accuracy of predicting whether someone needs a C-section out of all the women that eventually will (precision) would be guessing, which would be 50/50 with an identical recall of 50%. Running a logistic regression classification with all 28 predictors was only able to improve precision to 62% and dropped recall down to 31%, meaning that more than half of the women that end up needing a C-section would be missed which is worse than guessing.

Using a binary classification via logistic regression did not prove to be a powerful predictive method for determining if a woman would need a C-section. However the probability outputs of the classification demonstrated interesting findings. Using solely BMI as a predictor resulted in the logistic regression probabilities that were directly correlated with the average rate of C-sections at that level of BMI - at a BMI of 29.7, 30% of women underwent C-sections and the logistic regression probability was 30%. The logistic regression probability outputs are powerful predictors of *risk* of C-section, however when setting a threshold to definitively classify that risk, they will always be wrong some percentage of the time in relation to that risk unless it is 0% or 100%.

But overall, these findings demonstrate an interestingly close to perfect relationship between BMI and the *rate* of C-sections for women at that level of BMI, which can serve to improve forecasting for healthcare systems. Running a linear regression on BMI and rate of C-section results in an Adjusted R-Squared of 96.7% with a coefficient of 0.0111, meaning that a woman’s BMI is an almost perfect approximation of her risk of undergoing a C-section. Removing the extremes of underweight and overweight by looking at the section of the population with a BMI ranging from 18.5-50, the linear relationship’s Adjusted R-Squared value jumps to 99.9% with a coefficient of 0.0117 and a standard error of 0.00002.

Predicting if an individual woman will undergo a C-section is challenging, but predicting across groups of normal, overweight, and obese women on average can be accurately forecasted by simply multiplying each woman’s BMI by 117%.

**Infection Analysis**

One major complication that occurs during birth is infection. According to a study by Cynthia J. Berg on pregnancy related mortalities in the US, it is estimated that 10% of all pregnancy deaths are due to the onset of an infection. Infections are especially important to identify because they can often be masked by the general uncomfortableness and soreness that occurs during delivery. Using logistic regression we have attempted to find a way of predicting whether or not an infection will occur during birth.

The distribution among the 3,801,534 births in the kaggle dataset was 3,696,802 with no infection reported and 104,732 patients with at least one or more infections. The baseline prediction for this model is to assume all patients will not have an infection, which has an accuracy of 97.24%. Using Logistic regression and a decision tree

**References**

Al-Shawwa, Mohammed & Abu-Naser, Samy S. (2019). Predicting Birth Weight Using Artificial Neural Network. *International Journal of Academic Health and Medical Research (IJAHMR)* 3 (1):9-14.

<https://philpapers.org/rec/ALSPBW>

Bever AM, et al. Fetal growth patterns in pregnancies with first trimester bleeding: the NICHD Fetal Growth Studies. *Obstetrics and Gynecology*. 2018;

DOI: 10.1097/AOG.0000000000002616

<https://www.nichd.nih.gov/newsroom/news/050918-early-pregnancy-bleeding#>

CDC database:

<https://wonder.cdc.gov/wonder/help/Natality-expanded.html>

F.O. Dare, A.S. Ademowore, O.O. Ifaturoti, A. Nganwuchu,

The value of symphysio-fundal height/abdominal girth measurements in predicting fetal weight,

International Journal of Gynecology & Obstetrics, ISSN 0020-7292,

https://doi.org/10.1016/0020-7292(90)91018-L.

<https://www.sciencedirect.com/science/article/pii/002072929091018L>

Mayo Clinic - C-Sections:

<https://www.mayoclinic.org/tests-procedures/c-section/about/pac-20393655>

Berg CJ, Callaghan WM, Syverson C, Henderson Z. Pregnancy-related mortality in the United States, 1998 to 2005. Obstet Gynecol. 2010 Dec;116(6):1302-1309. doi: 10.1097/AOG.0b013e3181fdfb11. PMID: 21099595.

<https://pubmed.ncbi.nlm.nih.gov/21099595/>