Data Mining Final Project

Anuka Revi and Maria Gilbert

5/13/2021

Predicting Birth weight for Babies

Abstract

Using data from the National Bureau of Economic Research on births in 2018, we created a model that aims to predict a baby's weight at birth, based on information that would be known prior to the baby being born.

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. While the gender, time of birth, and name of the new baby are all exciting and great news for the entire family to have, pediatricians focus mostly on babies' development, growth, and weight. Estimating a baby's weight in particular is very important, since an unusually high birth weight can pose serious injuries to the baby and its mother, and an unusually low birth weight is associated with many pediatric health conditions.

Ultrasounds are often used to estimate a baby's weight prior to birth, 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 that I did not need to be induced. However, my daughter was born at over 8.9 lbs which was significantly larger than the estimated size, and it made my delivery more complicated. So while we were exploring project ideas, we came across the NBER natality birth data from 2018, which contains information about 50,000 births. Our goal is to predict the birth weight of a full-term baby (37+ weeks). If our model can be used by doctors to better estimate babies' weights at birth, this can make deliveries easier, reduce complications, and maybe even decrease the rate of maternal mortality. It is interesting to see how machine learning tools can help us estimate the birth weight.

Data Exploration

Our data set contains 50,000 randomly chosen samples from 2018 U.S. birth certificate data (https://www.nber.org/research/data/vital-statistics-natality-birth-data), with 42 variables: - X: a unique identifier

- birth_month: an integer from 1 to 12, indicating the month of the birth
- birth_time: an integer from 0 to 2359, indicating the time of birth
- sex: a factor with two levels, for male and female
- birth_weight: the birth weight (in grams). This is the variable we are hoping to estimate
- mother_age: an integer indicating the age (in years) of the mother
- mother_birthplace: a factor with three levels, indicating the mother was born inside the U.S., outside the U.S., or if this is unknown
- hospital: a factor with three levels, indicating whether the baby was born in a hospital or not, or if this is unknown
- mother_race: a factor with six levels, indicating whether the race of the mother is white, black, American Indian/Alaskan Native, Asian, Native Hawaiian and other Pacific Islander, or mixed race
- mother_maritalstatus: a factor with three levels, indicating whether the mother is married, unmarried, or if this is unknown
- mother_education: a factor with eight levels, indicating the level of education of the mother, categorized into 8th grade or less, 9th to 12th grade with no diploma, high school graduate or equivalent, some college but no degree, associate degree, bachelor's degree, master's degree, or doctorate/professional degree
- father_combinedage: an integer indicating the age (in years) of the father
- father_race: a factor with six levels, indicating whether the race of the father is white, black, American Indian/Alaskan Native, Asian, Native Hawaiian and other Pacific Islander, or mixed race
- father_education: a factor with eight levels, indicating the level of education of the father, categorized into 8th grade or less, 9th to 12th grade with no diploma, high school graduate or equivalent, some college but no degree, associate degree, bachelor's degree, master's degree, or doctorate/professional degree
- priorlive: an integer indicating the number of prior live births the mother has had

- priorterm: an integer indicating the number of prior term births the mother has had
- time_sincelastbirth: a factor indicating whether the mother's last birth occurred less than 18 months before the current birth, 18 months to under 2 years before the current birth, 2 years to under 4 years before the current birth, 4 years to under 6 years before the current birth, greater than 6 years before the current birth, or whether this is N/A or unknown
- prenatal_visits: an integer indicating the number of prenatal care visits
- wic: a factor indicating whether the mother participates in WIC (Special Supplemental Nutrition Program for Women, Infants, and Children) or not, or whether this is unknown
- f_cigs_1: a factor indicating whether the mother smoked during the first trimester
- f_cigs_2: a factor indicating whether the mother smoked during the second trimester
- f_cigs_3: a factor indicating whether the mother smoked during the third trimester
- m_ht_in: an integer indicating the height (in inches) of the mother
- bmi: the BMI (Body Mass Index) of the mother
- prepreg_weight: an integer indicating the weight (in pounds) of the mother before the current pregnancy
- wtgain: an integer indicating the amount of weight (in pounds) the mother gained throughout the pregnancy
- gest_diabetes: a factor indicating whether the mother has gestational diabetes or not, or if this is unknown
- hypertension_eclampsia: a factor indicating whether the mother has hypertension eclampsia, or if this is unknown
- infertility_treatment: a factor indicating whether the mother underwent fertility treatment for the current pregnancy, or if this is unknown
- previous_cesarian: a factor indicating whether the mother has previously given birth via Cesarian section, or if this is unknown
- no_infections: a factor indicating whether the mother has any infections
- labor_induced: a factor indicating whether the labor was induced or not, or if this is unknown

- anesthesia: a factor indicating whether the mother was given anesthesia for the birth or not, or whether this is unknown

- delivery_method: a factor indicating whether the birth was vaginal, C-section, or if this is unknown

- ruptured_uterus: a factor indicating whether the mother experienced uterine rupture during birth or not, or if this is unknown

- perineal_laceration: a factor indicating whether the mother experienced perineal lacerations (known commonly as vaginal tears) during the birth or not, or if this is unknown

- attend: a factor indicating whether the birth was attended by a medical doctor, a midwife, or if this is unknown

- apgar5: a integer indicating the baby's APGAR (Appearance, Pulse, Grimace, Acitivty, and Respiration) score five minutes after birth. This score is used as a summary of the baby's overall health and is used to determine whether the baby needs medical attention. It is an integer ranging from 0 to 10, with 0 being an emergency and 10 being ideal

- plurality: a factor indicating whether or not the birth was a multiple birth

- combined_gestation: an integer indicating which week of gestation the baby was born at

- breastfed: a factor indicating whether the baby was breastfed after birth or not, or if this is unknown

Many of these variables will not be relevant in our models, because of the way that they might be caused by the birth weight, or emerge after the birth weight would already be known. We will be excluding breastfed and apgar5 because these emerge after the birth has already occurred. We will also be excluding perineal_laceration, ruptured_uterus, delivery_method, and hypertension_eclampsia, since these emerge during birth, and are influenced by the weight of the baby. Finally, we will be excluding all data on plural births, as well as pre-term births. This way we can control for the fact that pre-term births nearly always have low birth weight.

In our initial exploration of the data, we evaluated the correlations between all variables, and examined the variance in values for some selected variables. Figure 1 shows the correlations between our variables.

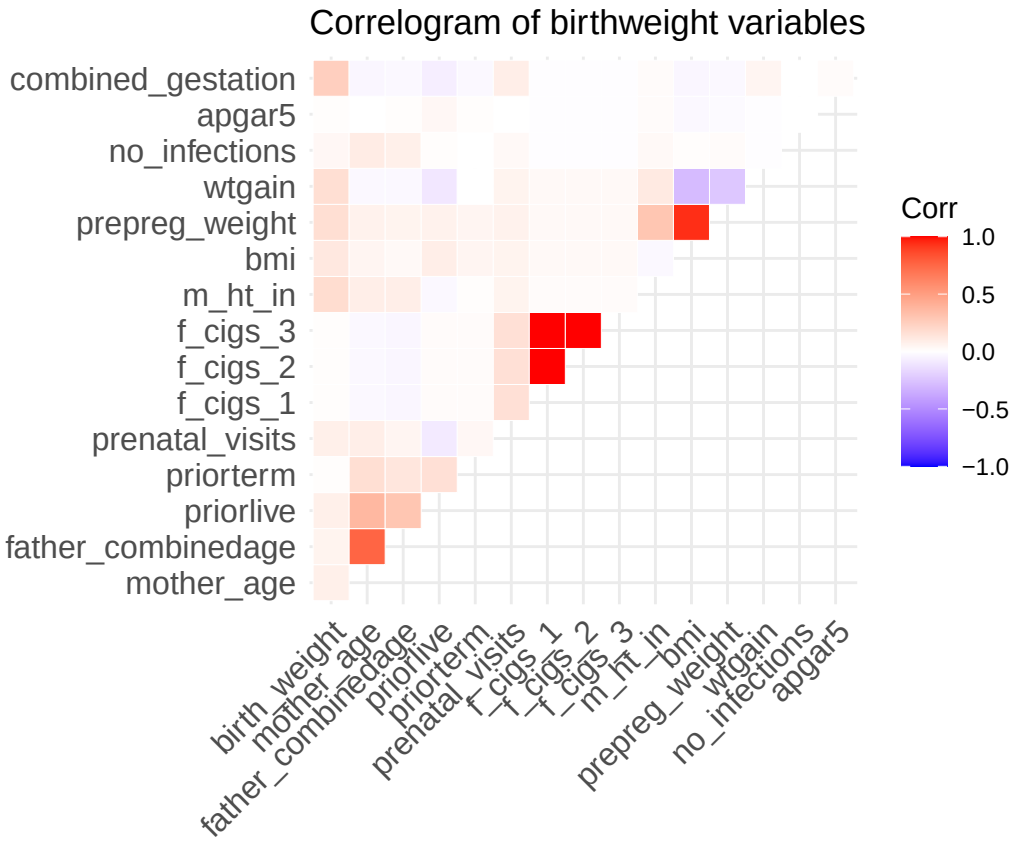

Figure 1: Correlogram

As we can see, f_cigs_1, f_cigs_2, and f_cigs_3 all have the same correlations with the other variables. The strongest correlations with birth weight are weeks of gestation (even with pre-term births filtered out!), weight gain, pre-pregnancy weight, BMI of the mother, and the height of the mother, all having a positive correlation with birth weight.

Next, we aimed to examine the distributions of some of our variables. Figure 2 shows the distribution of actual birth weights from our data.

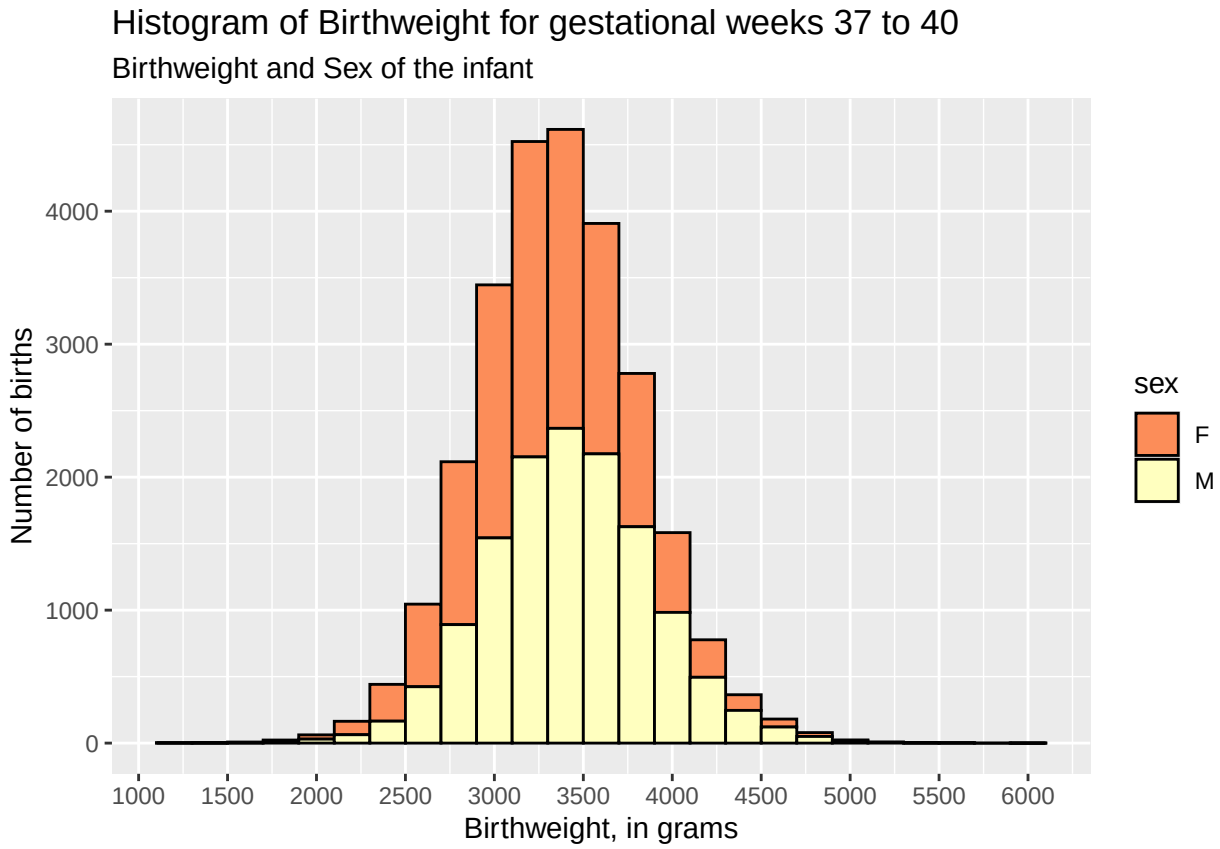

Figure 2: Birth weight Distribution

Figure 3 shows distributions of selected variables within our data relating to the mother's medical history.

Figure 4 shows distributions of selected variables within our data relating to the mother's demographics.

The x-axis labels for Mother Education Level have been abbreviated for readability. - "1-8" = up to 8th grade - "some" = some high school but no diploma - "HS" = high school diploma or equivalent (e.g., GED) - "some" = some college but no degree - "Assoc" = Associate degree - "Bach" = Bachelor's degree - "Mast" = Master's degree - "PhD" = Doctorate (PhD, EdD) or professional degree

Now, we want to look at the relationships between some of these variables and the birth weight of the baby. Figure 5 shows some of these relationships, with LOESS (local regression) trend lines. The vertical black lines show the mean values within our data of weeks of gestation, weight gain, pre-pregnancy weight, and height.

Prior to analysis, we converted categorical variables into factors and re-coded them in accordance with the code book provided by NBER. We replaced unknown variables that were coded as 99, 999, or 9999 with NA values and renamed variables more intuitively for readers. Since we were interested in the birth weight outcome, we decided to filter the data and focus on babies who were born on 37 to 40 weeks mark, since there is a strong causal relationship between week of gestation and the potential birth weight outcome. As we would expect, babies who are born at 37-40 weeks weigh more than premature babies.

In our processed data *week37to40* we have 2,134 missing values out of the total 1,159,808 observed values. The variable mother_race contains 2,061 missing values. Since these NA values are close to 10% of our 26,162 observations, we had to ignore the mother's race in our analysis. Additionally, we had eliminated time_sincelastbirth due to having mostly NA values in the original data set.

Methods and Results

We first used a random forest model to evaluate variable importance.

Although the correlations provided a comprehensive overview of the most important numeric variables and multicollinearity among them, we also used a random forest model to get an overview that would include the categorical variables. A random forest model works by creating a multitude of decision trees and selecting the mean training value at each node. This method allows us to find the variable importance without much effort or parameter tuning. As we would expect, some of the most important variables are combined_gestation, wtgain, prepreg_weight, mother's bmi, height, race, age, prior live births, and sex of the baby. Figure 6 shows the variable importance plot that resulted from our random forest.

The RMSE of this random forest model is:

```
## [1] "RMSE =  408.201242988692"
```

We were interested in creating models that would predict two different outcomes. One would be the birth weight in grams, and another would be whether the baby was born with a low weight (under 2,500 grams or 5.5 pounds), a high weight (over 4,000 grams or 8.8 pounds), or a normal weight (2,500 to 4,000 grams).

**Part A: Models to Predict Whether Baby Will be Above or Below Average Weight

In order to predict whether a baby would be born with a low weight, normal weight, or high weight, we created two xgboost models. The first determined whether the baby would be low weight or not, and the second determined whether the baby would be high weight or not. The following shows the feature importance plot of the first xgboost model, which predicts whether or not a baby will be low weight. We calculated the test error of this model to be:

```
## [1] "Test error =  0.0255351681957187"
```

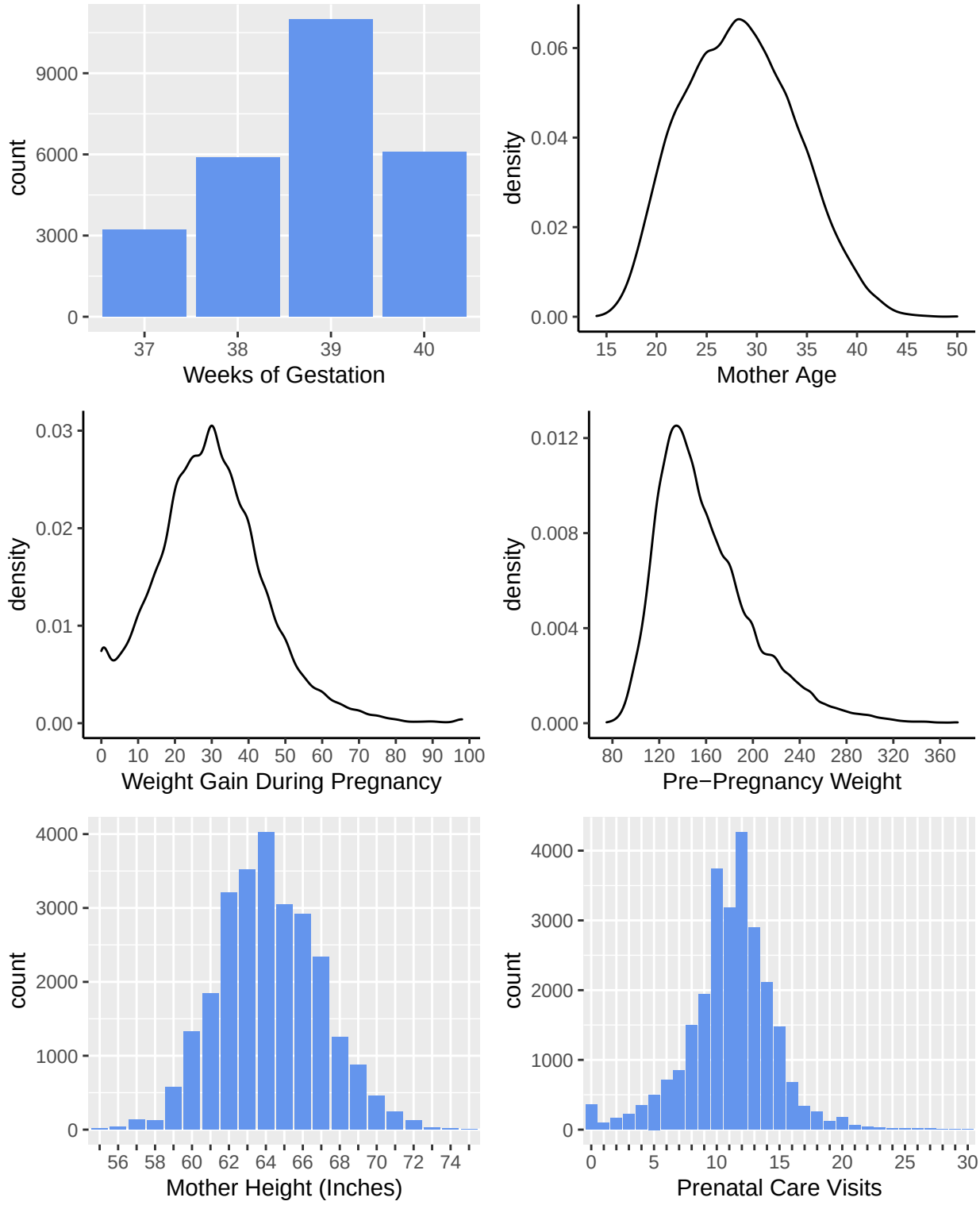

Figure 3: Maternal health history distributions

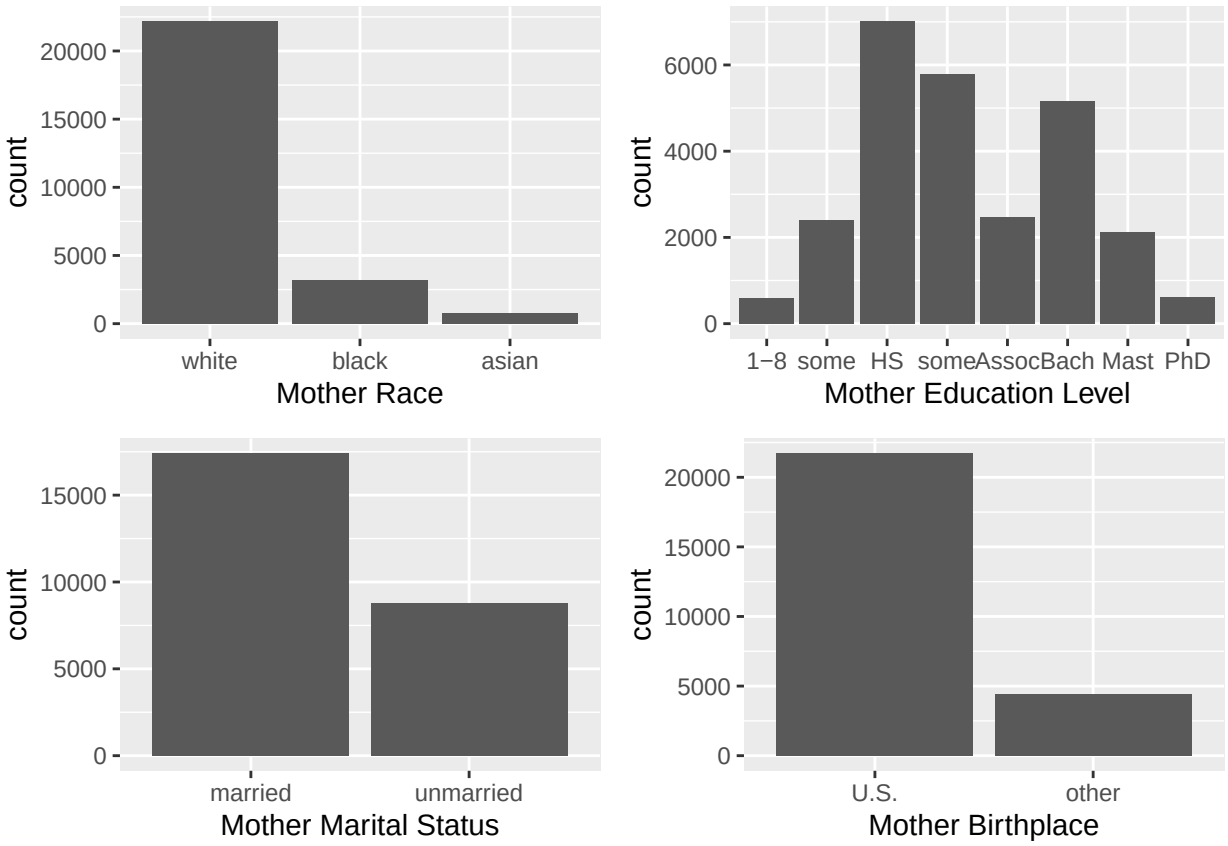

Figure 4: Maternal demographic distributions

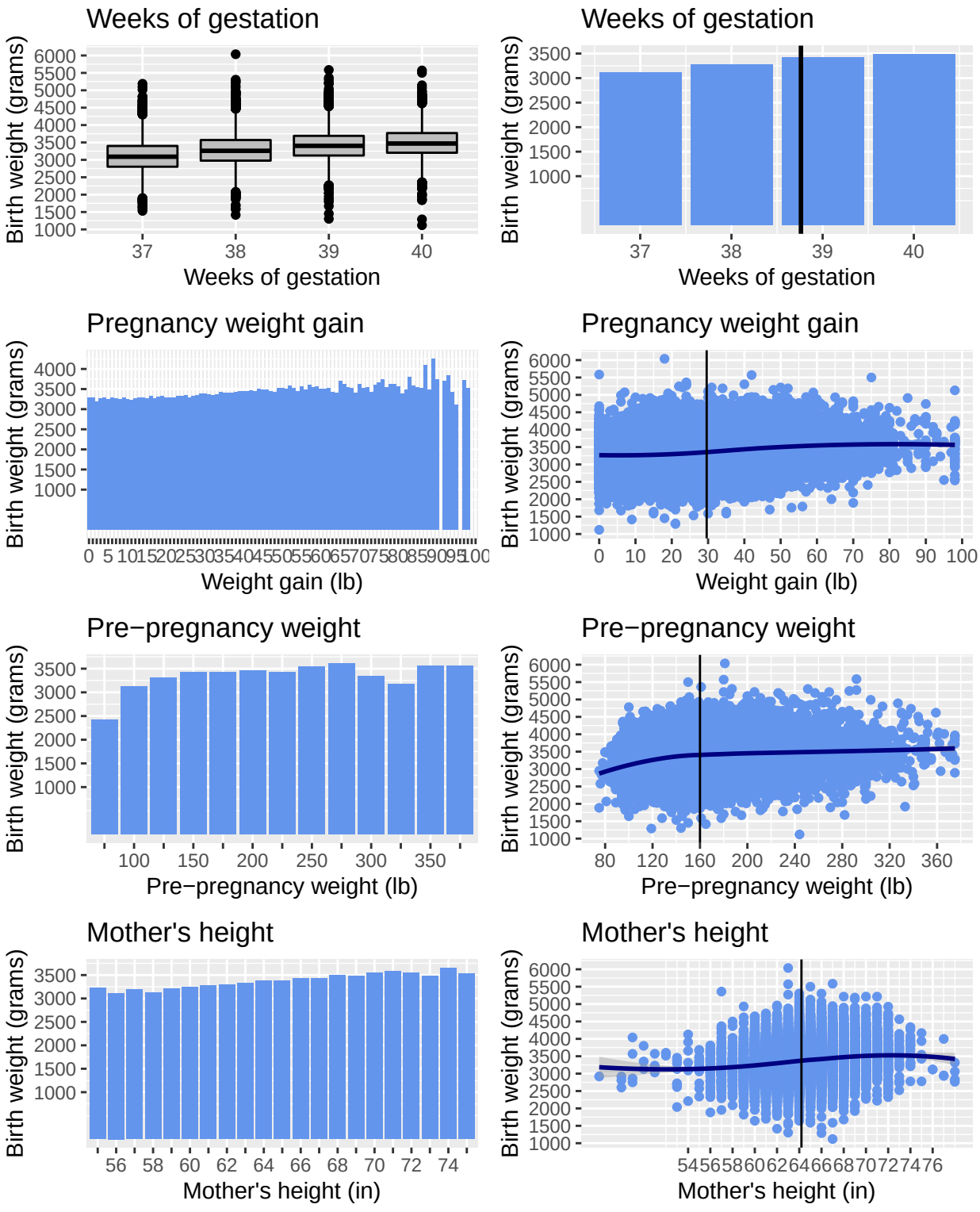

Figure 5: Maternal health history and birth weight

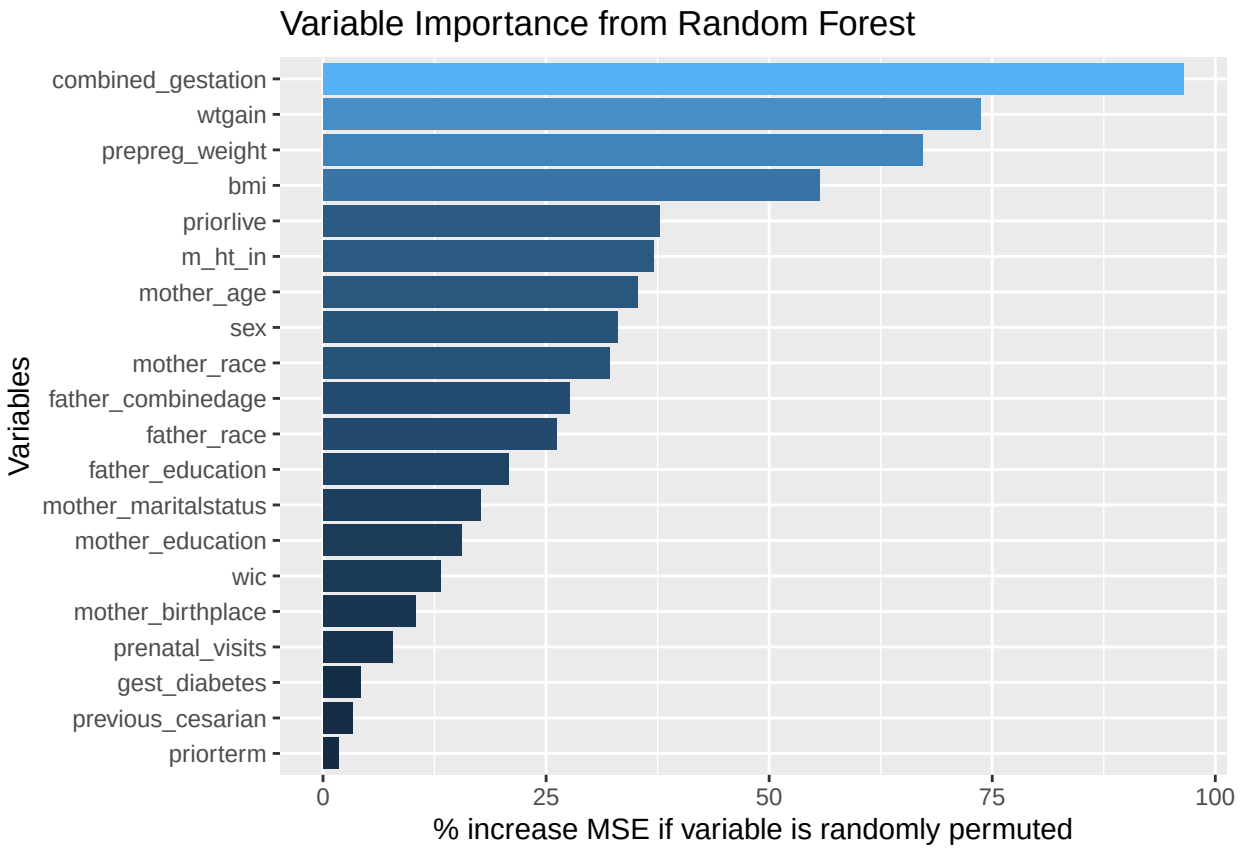

Figure 6: Variable Importance Plot

This means that about 97.5% of the time, this model is correct in predicting whether or not a baby will be born with a low weight. Figure 7 shows the feature importance of this model.

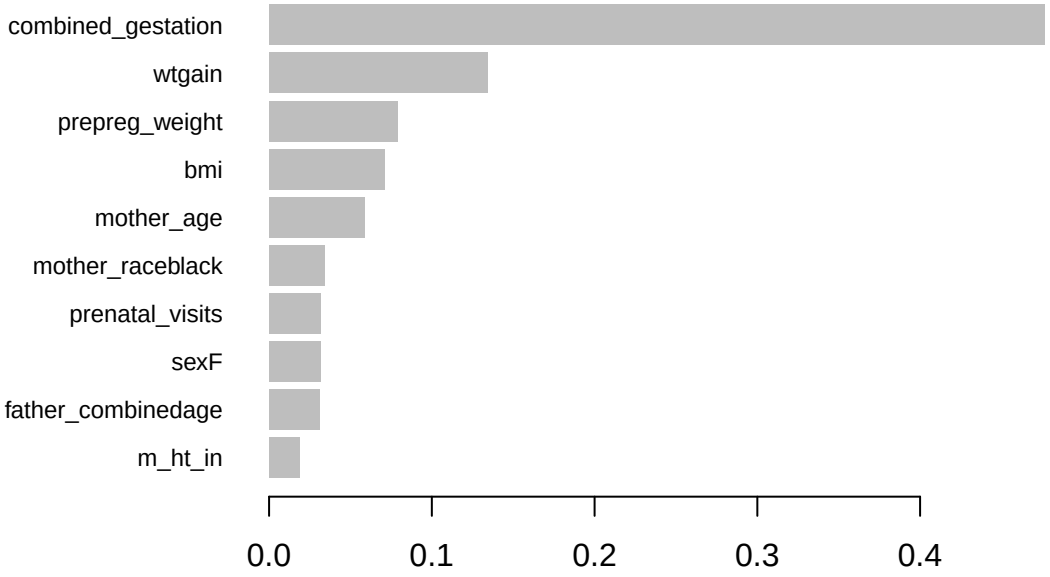

Figure 7: Feature importance for low birth weight

```
##                Feature       Gain       Cover  Frequency Importance
##  1: combined_gestation 0.47656694 0.485118110 0.15107914 0.47656694
##  2:             wtgain 0.13450278 0.166447001 0.20863309 0.13450278
##  3:     prepreg_weight 0.07881839 0.116686405 0.14388489 0.07881839
##  4:                bmi 0.07118682 0.068074787 0.10791367 0.07118682
##  5:         mother_age 0.05872167 0.050693153 0.09352518 0.05872167
##  6:   mother_raceblack 0.03432288 0.009720851 0.02158273 0.03432288
##  7:    prenatal_visits 0.03177786 0.023080345 0.06474820 0.03177786
##  8:               sexF 0.03171858 0.029585526 0.02877698 0.03171858
##  9: father_combinedage 0.03130925 0.029814649 0.07194245 0.03130925
## 10:            m_ht_in 0.01867703 0.004615495 0.02158273 0.01867703
```

We created a similar model to predict if baby will be born with a high weight, and calculated the test error of this model to be:

```
## [1] "Test Error =  0.0830275229357798"
```

This means that about 91.7% of the time, this model is correct in predicting whether or not a baby will be born with a high weight. Figure 8 shows the feature importance of this model.

```
##                Feature        Gain       Cover  Frequency  Importance
##  1:     prepreg_weight 0.310133847 0.366469487 0.17687075 0.310133847
##  2:             wtgain 0.277613573 0.254670402 0.25850340 0.277613573
##  3: combined_gestation 0.131629444 0.143218862 0.17687075 0.131629444
##  4:               sexF 0.093856125 0.084674986 0.10204082 0.093856125
##  5:          priorlive 0.061552796 0.044200680 0.06802721 0.061552796
##  6:   mother_raceblack 0.039844964 0.036247390 0.03401361 0.039844964
##  7:            m_ht_in 0.028459579 0.019127682 0.02721088 0.028459579
##  8: father_combinedage 0.020487275 0.014856487 0.03401361 0.020487275
##  9:     gest_diabetesY 0.014823729 0.010016019 0.02040816 0.014823729
## 10:         mother_age 0.007995638 0.008758393 0.03401361 0.007995638
```

While weeks of gestation is the most important feature in telling whether the baby will be born with low

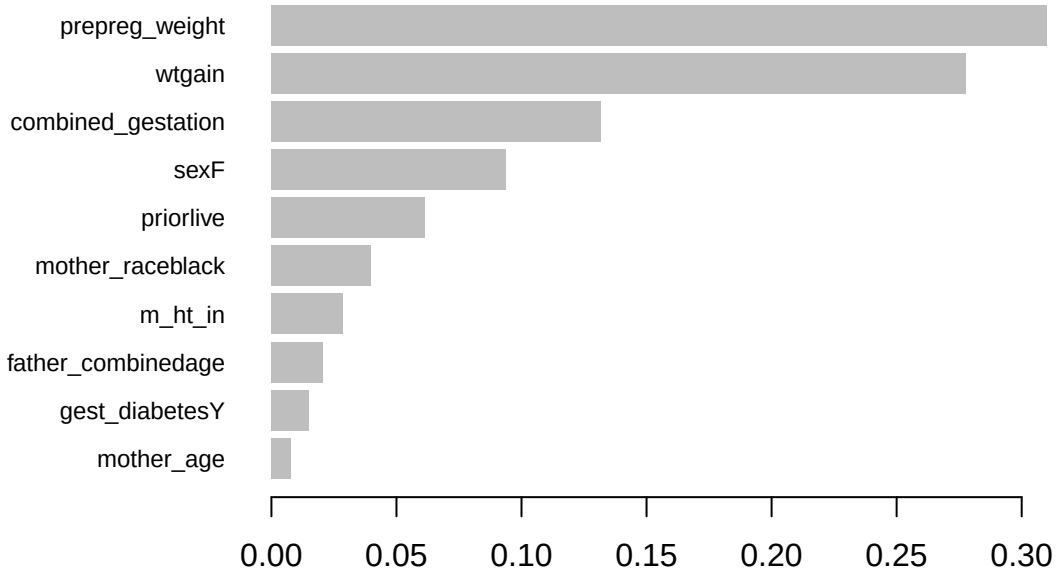

Figure 8: Feature importance for high birth weight

weight (even considering only full term births), pre-pregnancy weight and pregnancy weight gain are the most important in telling whether the baby will be overweight. We can tell from this that there are some factors that are more important at the lower end of the spectrum of birth weight than those that are more important at the higher end of the spectrum.

Part B: Predicting Actual Birth Weight in Grams

Now, our last model aims to predict the actual birth weight in grams. We tried several different types of models, but the two with the highest accuracy were lasso and step, with their accuracy levels being virtually identical. A lasso (Least Absolute Shrinkage and Selection Operator) model is from an algorithm that performs variable selection and regularization and then results in a regression model. A step model is from an algorithm that starts with a basic linear regression model and tests iterations with different interactions between the variables until it finds the optimal linear model. Because these models are similar in how they work, it is not surprising that we found similar RMSE (root mean squared error) in both of them.

```
## [1] "best lambda value is 0.43044093571137"
```

```
## [1] "RMSE =  407.363324349081"
```

The relative error improvement of our final lasso model over the random tree that we used to examine variable importance was

```
## [1] 0.00205271
```

Not a significant improvement, but both models do a decent job at predicting the birth weight given the information available before the birth.

Conclusion

We found that using only variables that would be known before a baby is born, it is possible to predict whether a baby will be low weight, normal weight, or high weight, and even to predict the actual numeric weight itself, all with modest to reasonable accuracy. This model could be used by doctors to supplement

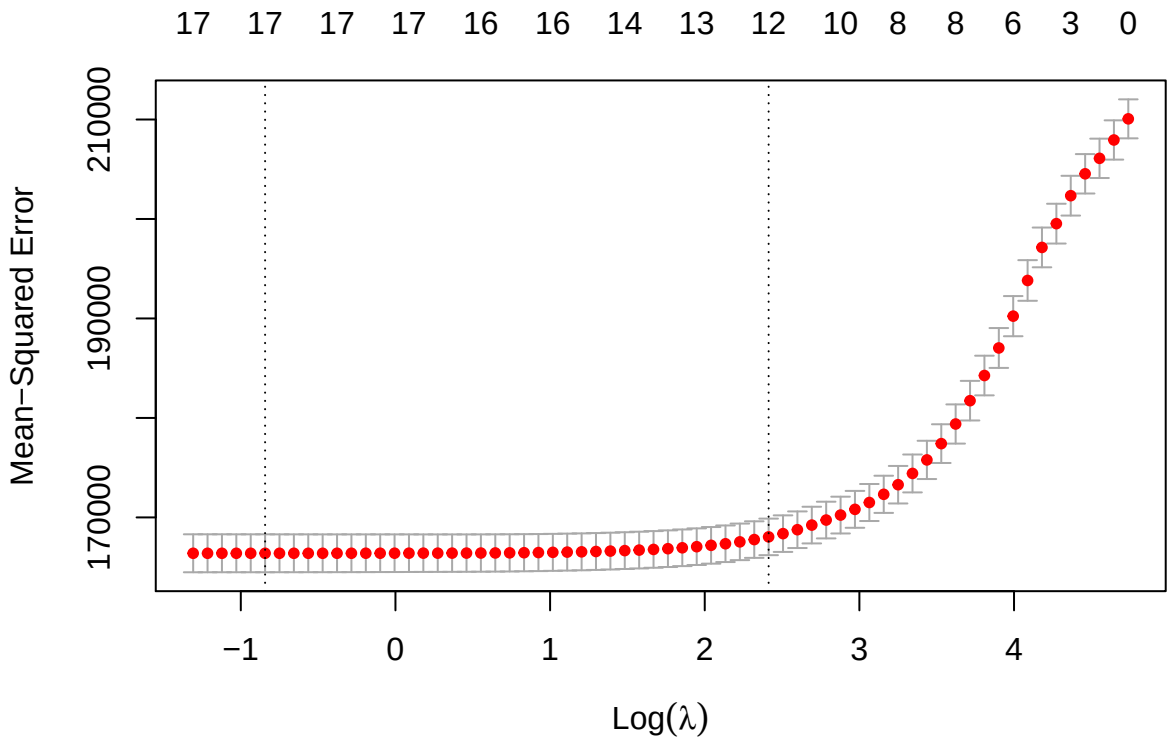

Figure 9: Lambda and MSE

fetal weights estimated by ultrasound to get a more accurate guess for how much a baby will weigh at birth, and as a result, how likely it will be for the mother and/or baby to have complications during birth.