Birthweight Prediction

Modeling the relationship between maternal attributes, behaviors and birthweight

# Executive Summary:

It is an unfortunate yet accepted fact that babies born at a “low” birthweight tend to have complications not only postpartum, but also potentially throughout life. Early on from breathing, digestive and immune system issues, to later on in life in the form of diabetes, heart disease, and high blood pressure – these all cause a lifetime of challenges and can influence an individual’s quality of life (March of Dimes, 2018). Therefore, understanding the potential factors that lead to low birthweight is important from multiple perspectives. The first, and objectively the most important, is the mother’s – knowing and understanding the risk factors that lead to low birthweight may provide guidance and potentially alter behavior to lessen the risk of having a baby born with low birthweight. Secondly, from a health care provider perspective, these factors can provide not only educational opportunities for expecting mothers but also allow for the planning of proper resources postpartum. Finally, from a public health perspective, the understanding of these factors can guide further areas of research into the subject and help determine how and where to allocate resources.

In this study, we attempt to model birthweight (measured in kg) using eight (8) different variables, which can be classified as either as behaviorally related (smoker, weight, etc.), demographically related (age, race, etc.), or health related (presence of hypertension, uterine irritability, etc.). We wish to model the outcome as either “low” birthweight (i.e. <2.5kg) or “not low” birthweight (>2.5kg), which is known as a binary response. We choose this binary response in the data set since it agrees with World Health Organization standards (WHO, 2014), and therefore provides a good threshold to flag higher risk vs. lower risk given all the aforementioned reasons.

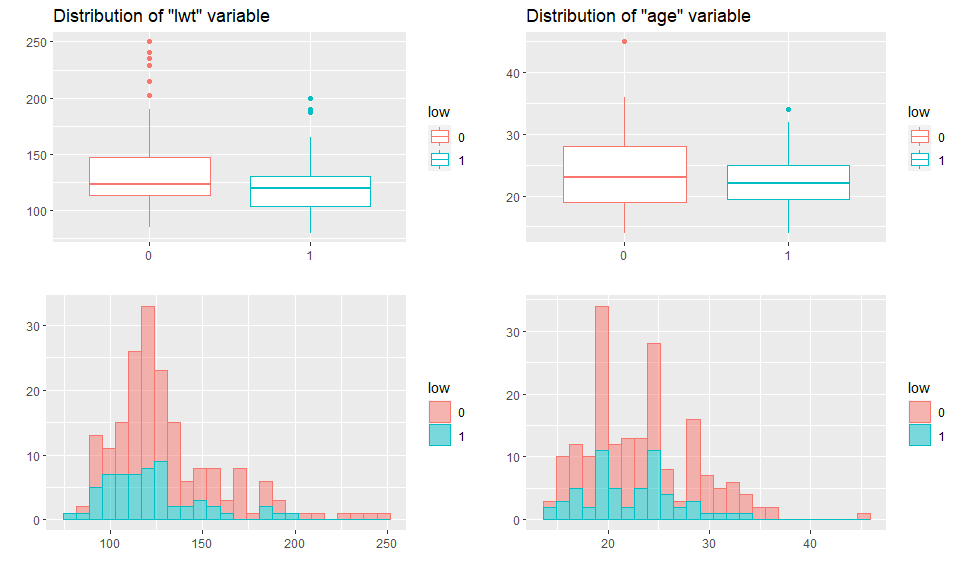
To get us started, we begin our search for data mining techniques that are used to predict classification-based outcomes. While there are many options and tools for the job, we must consider pros and cons of these various techniques, as well as check the mathematical assumptions that must be met regarding the dataset. We start with the summary statistics of our dataset and visualize distributions and relationships among variables. A few things to note upon initial inspection: there are no data with missing values, therefore no omission or imputation of observations are performed. The quantitative variable lwt (Mother’s weight) has a distribution with a handful of outliers along both response classes, and appears skewed to the right. A log transformation on the lwt predictor will be performed to account for the skew. The quantitative variable Age has a distribution with a only one outlier along both classes (Fig. 1). Finally, we check correlation between predictors, however, there doesn’t appear to be any strong linear relationships amongst them (no concern with multicollinearity).

Figure 1: Variable Distribution

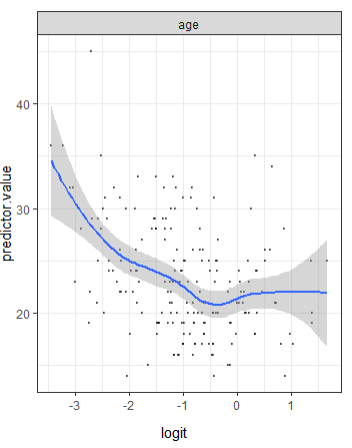
Now that we have a bit more of an understanding of our data set, we match up these characteristics to potential modeling techniques - we begin by exploring the well-known logistic regression, which is a “transparent” method that allows us insight into which individual predictors influence the response, and by how much. However, upon further inspection it does not meet all of the assumptions (specifically linearity along the logit for the Age variable – fig. 2). We also check assumptions for linear/quadradic discriminate analysis; however, the assumptions (multivariate normality per class checked) of these methods are not met either. Given that we don’t have much linearity among predictors and the response, and we do have a handful of outliers, we opt for more flexible models such as the K-nearest neighbors (does not require linearity), and the Random Forest (data can be non-linear, and isn’t highly suspectable to outliers since it chooses subsets of observations/predictors at random during its fitting process).

Figure 2: Non-linearity along the logit

We begin our model selection process by fitting our K-nearest neighbors (KNN) model to the prepared data. The unique challenge with KNN is two-fold – first we must standardize our quantitative predictors (so they are on the same scale), and then we must find the optimal value for a tuning parameter (K). We use 10-fold cross-validation to solve for this challenge, and identify the model that minimizes misclassification error rate as K=1 and all predictors except mother’s weight, at 14.28% (fig. 1).

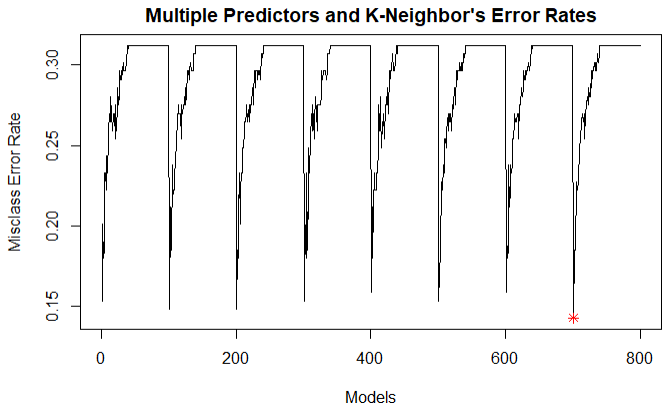


Figure 3: \* icon denotes best KNN model

We now move onto our second model of choice, the random forest, which also has a tuning parameter like KNN (“mtry”,which limits the number of parameters allowed for each tree fitted). We also do predictor selection here by adding one predictor at a time . We take a similar approach of 10-fold cross-validation to identify our tuning parameter/predictors and find that a random forest model with 5 predictors (age, race, smoke, ptl, ht) and an mtry=1, results in the best model with an error of 29.10% (fig. 4).

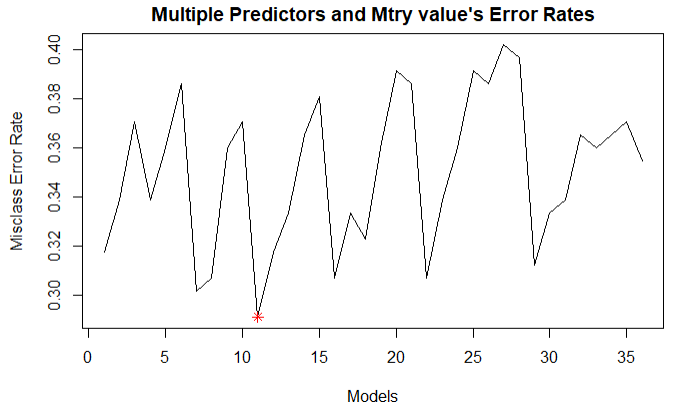


Figure 4: \* icon denotes best Random Forest model

Overall, it appears the KNN model with all predictors except lwt (with K=1) is superior to the Random Forest, and is therefore our top choice in this selection process.

Finally, to get an honest assessment of performance for our top model (i.e. how well will it generalize and perform with honestly new data), we complete a double cross-validation using the modeling selection process as discussed above. We choose this technique because the overall sample size of the data set (189 observations) is too small to justify a single train/test split. The output of this honest assessment process shows our KNN model (K=1, 7 predictors) can be expected to perform with an overall misclassification error rate of 29.62% on truly new data.

In conclusion, we opt for a less interpretable model due to the nature of the data available. This model allows us to fit our data well with an accuracy of ~70% on future predictions. Although we also discover that age, race, smoking status, number of previous preterm labors, having hypertension, uterine irritability, and how many times the mother visits the physician in the first trimester are most important, we are unfortunately not able to tell by “how much” each of these variables influences the overall likely outcome.

References:

March of Dimes. (2018, March). *Low Birthweight*. <https://www.marchofdimes.org/complications/low-birthweight.aspx>

World Health Organization. (2014). *Global Nutrition Targets 2025: Low birth weight policy brief.* <https://www.who.int/nutrition/publications/globaltargets2025_policybrief_lbw/en/>