STAT479 Project II: Prediction of Low Birth Weight

Wenyan Zhou

1 Introduction

What if a baby is born at low weight without any preparation? What factors lead to low birth weight? it is necessary to make extra preparations in case a baby' s weight is lower than normal, and know the possible reasons. The goal of this project is to predict whether a baby is born at low weight given training data containing related variables from 2011 to 2016.

2 Data description & preparation

The project data is from a longitudinal survey sponsored by the National Bureau of Economic Research from 2011 to 2016. It collects information on a baby' s birth weight as well as characteristics of its parents and other conditions before birth. The whole data set has more than twenty million records. To avoid stretch computer resources, I use a representative subset of it. First, I delete those with "dbwt" missing, where "dbwt", represents a baby' s birth weight in gram. Then, I create a dependent variable lowbwt = I (dbwt < 2500), the variable we are interested in, as an indicator of low birth weight. It shows that about 92% babies are born at normal weight, and only 8% are born at low birth weight. To make the structure of the subset closer to the whole data set as well as the given test sample, I do the following steps:

1. Select 112 independent variables used in the test sample along with lowbwt;
2. Determine each variable' s type and perform some conversion and encoding to meet requirements;
3. Separate each year' s data into two parts: (a) lowbwt = 0 and (b) lowbwt = 1;
4. Randomly sample 113160 from each year' s (a) in step 2, and 9840 from each year' s (b), both with no replacement;
5. Divide the 12 parts obtained in step 3, and combine in row into 6 small training subsets, each containing 120,000 records;
6. Additionally, use the rest of the data to create a validation set the same size as the test

sample for model evaluation;

After preparations, I get a stratified clean sample with a reasonable size, both for accuracy and efficiency. Actually, I tried to use almost all observations for one model fitting in advance, but it was quite time consuming and didn't give me a significant lower error rate.

To determine the type of a variable in step 2, whether it should be considered as categorical or numerical, some variables are ambiguous. For example, "feduc", representing father's education, is divided into 9 levels in Table 1:

Table 1: Details of variable "feduc"

1	8^{th} grade or less
2	9^{th} through 12^{th} grade with no diploma
3	High school graduate or GED completed
4	Some college credit, but not a degree
5	Associate degree (AA, AS)
6	Bachelor's degree (BA, AB, BS)
7	Master's degree (MA, MS, MEng, MEd, MSW, MBA)
8	Doctorate (PhD, EdD) or professional Degree (MD, DDS, DVM, LLB, JD)
9	Unknown

It seems that this variable can be regarded as a numerical variable as it is in an increasing trend. However, use the validation set for plot, we can see from Figure 1, color blocks are not in a gradient order, which means there is no obvious correlation between "feduc" and "lowbwt".

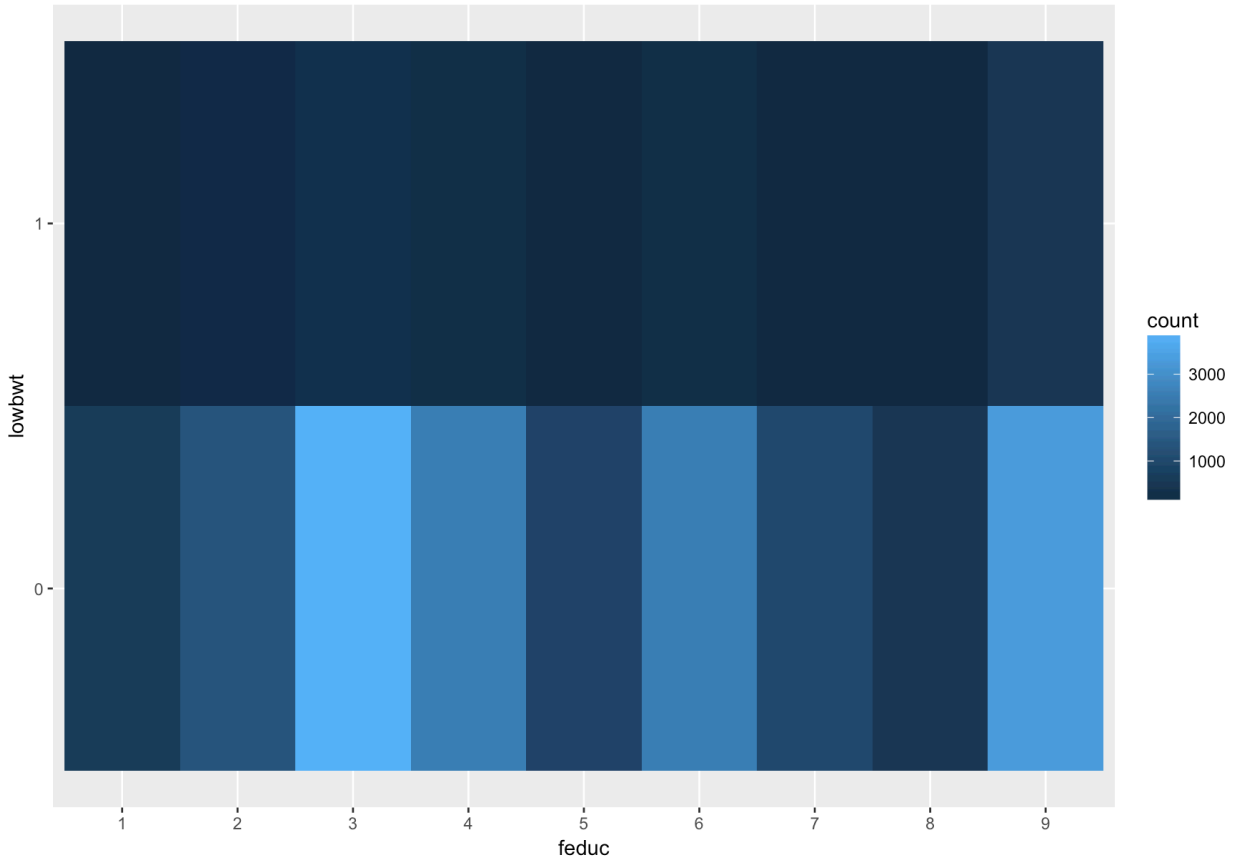


Figure 1: lowbwt vs. feduc

It may cause some problems for linear models. Therefore, for variables like "feduc", I classify them as categorical.

As a result, I classify variables "bmi", "cig_0", "cig_1", "cig_2", "cig_3", "combgest", "dwgt_r", "fagecomb", "illb_r", "ilop_r", "m_ht_in", "mager", "priorterm", "pwgt_r", "rf_cesarn", "wtgain" as numerical, while others categorical.

3 Methods for model fitting

Five methods are tried for prediction, three of them are GUIDE tree-based methods, one is linear model and the other is an ensemble method using both boosting trees and linear boosting.

GUIDE classification tree. This is the very basic GUIDE tree method.

GUIDE forest. This is an ensemble of 500 unpruned GUIDE trees constructed by the GUIDE

classification tree method without interaction and linear splits. As in random forest, GUIDE forest uses a random subset of $\sqrt{K}$ variables to split each node.

GUIDE bagging. This is an ensemble of 100 pruned GUIDE trees, each constructed using the GUIDE classification tree method from a bootstrap sample.

Logistic regression. The model is:

$$\mathrm{P}(\mathrm{Y}=1|\mathrm{X}=\mathrm{x})=\frac{e^{x\prime\beta}}{1+e^{x\prime\beta}}.$$

Xgboost. This is an ensemble method short for is short for eXtreme Gradient Boosting. It is an efficient and scalable implementation of gradient boosting framework. Two solvers are included: CART tree learning algorithm and linear model.

4 Application to data

Before feed data to models, we should consider two issues beforehand, one is missing values, and the other is extremely unbalanced response.

For missing values, when running GUIDE, it can impute missing values automatically, so I do not need to do imputation beforehand. However, for other machine learning methods, they will crash with missing values. Missing values in categorical variables are already encoded as "U" in data processing. Therefore, I use GUIDE regression tree to impute missing values in numerical variables. Figure 2 shows missing values in numerical variables. We can see that variable "ilop_r" has more than 80% missing.

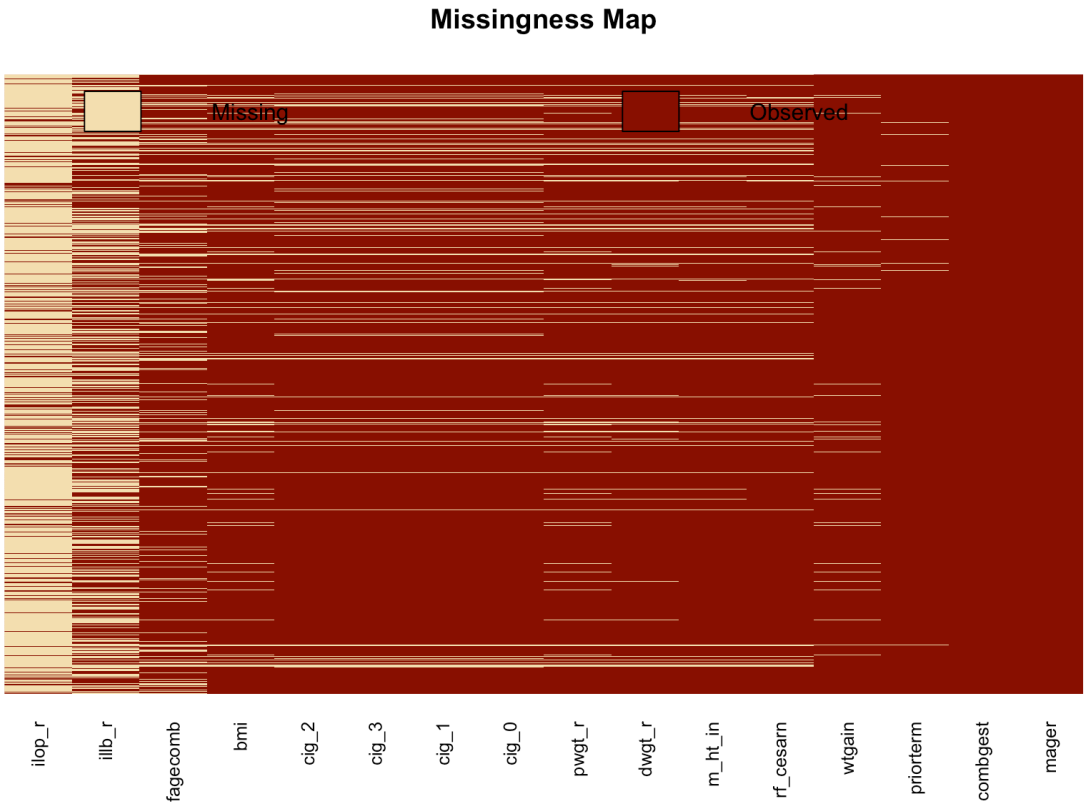


Figure 2: missing values vs. observed

For unbalanced response, I use a 10:1 cost matrix:

		True lowbwt	
		0	1
Predicted lowbwt	0	0	10
	1	1	0

Regarding this cost matrix,

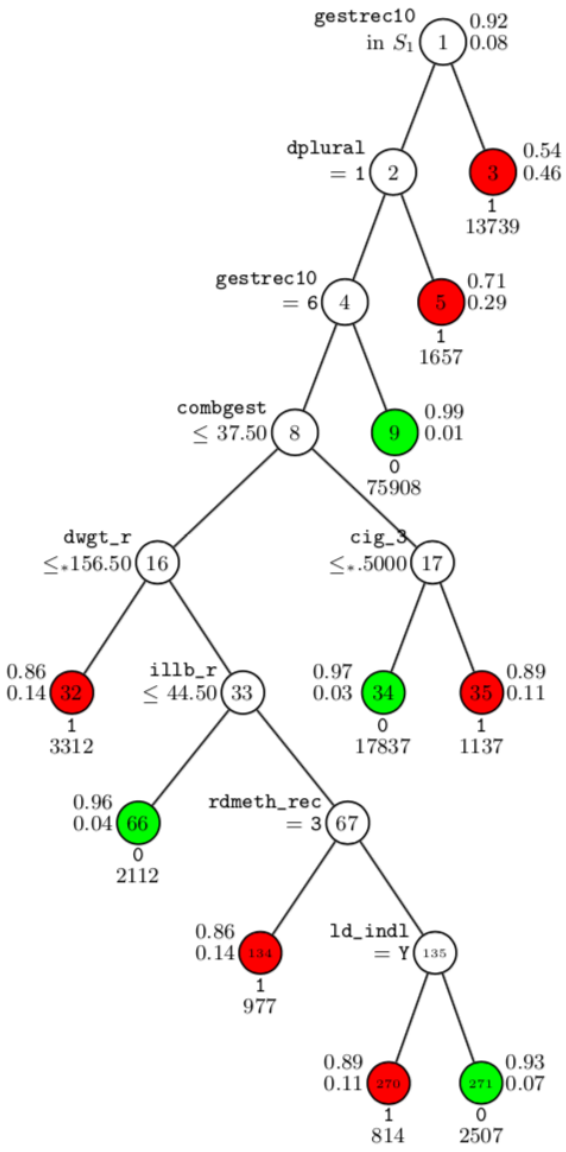
$$error\ rate = \frac{\#\ False\ Positive * 10 + \#\ False\ Negative}{\#\ records}$$

4.1 GUIDE classification tree

Table 2: Error rate on each training subset and on validation set (GUIDE classification tree)

Data	Error rate
Subset1	0.28455
Subset2	0.2842833
Subset3	0.285975
Subset4	0.289525
Subset5	0.2867583
Subset6	0.27865
Validation set	0.2895

Table 2 shows error rate on each subset using GUIDE classification tree. For validation set, I use six models obtained from six subsets to make predictions and vote for the final prediction. It is also used for the other two GUIDE methods.



GUIDE v.27.2 0.50-SE classification tree for predicting `lowbwt` using estimated priors and specified misclassification costs. At each split, an observation goes to the left branch if and only if the condition is satisfied. The symbol '$\leq_*$' stands for '$\leq$ or missing'. Set $S_1 = \{10, 6, 7, 8, 9\}$. Predicted classes and sample sizes printed below terminal nodes; class proportions for `lowbwt` = 0 and 1 beside nodes. Second best split variable at root node is `gestrec3`.

Figure 3: GUIDE classification tree on subset1

Figure 3 gives some information about tree constructing process based on subset1. I use it as an example because the other five trees are almost the same. We can see clearly that variables like "gestrec10", "dplural", and "combgest" are the most important ones,

representing combined gestation recode 10, plurality recode, and combined gestation-details in week respectively. This result really makes sense because premature infants have no enough time for growing, and twins or more infants having the same mother need to share nutrition. It is quite reasonable that they born at low birth weight. Variables like "dwgt_r" and "cig_3" are also important, which gives pregnant women a warning not to lose weight and smoke during pregnancy period for the sake of their babies.

4.2 GUIDE forest

Table 3: Error rate on each training subset and on validation set (GUIDE forest)

Data	Error rate
Subset1	0.2765917
Subset2	0.2779583
Subset3	0.2792
Subset4	0.2813
Subset5	0.2768583
Subset6	0.2704917
Validation set	0.2750556

Table 3 shows error rate on each subset using GUIDE forest. Overall, it's error rates are slightly lower than GUIDE classification tree by 1%.

4.3 GUIDE bagging

Table 4: Error rate on each training subset and on validation set (GUIDE bagging)

Data	Error rate
Subset1	0.2789444
Subset2	0.2873333
Subset3	0.2894444
Subset4	0.2875

Subset5	0.2833333
Subset6	0.2916667
Validation set	0.2893889

Table 4 shows error rate on each subset using GUIDE bagging. It is almost the same as single tree method.

4.4 Logistic regression

Logistic regression will give an error if the validation set contains class values that do not appear in the training sample. To avoid this, I use the combination of six subsets for model fitting, and 0.08 as a threshold to classify. This method gives me an error rate of 0.2698333 on the validation set. It is lower, but not comparable with GUIDE methods as they are not fed with exactly the same data.

Figure 4 is part of R output of the model. I rank variables in an increasing order of p-value.

	Estimate	Std. Error	z value	Pr(>|z|)
combgest	-5.214983e-01	9.503816e-03	-5.487252e+01	0.000000e+00
ld_sterY	1.363689e+00	3.131822e-02	4.354297e+01	0.000000e+00
dplural2	2.258011e+00	6.732187e-02	3.354052e+01	1.237595e-246
sexM	-3.837733e-01	1.160915e-02	-3.305784e+01	1.200228e-239
ld_indlY	3.505980e-01	1.595869e-02	2.196909e+01	5.689360e-107
rf_pptermY	6.018225e-01	2.823287e-02	2.131638e+01	8.001748e-101
mtranY	1.155623e+00	5.890536e-02	1.961831e+01	1.078883e-85
dplural3	4.273822e+00	2.390161e-01	1.788090e+01	1.661401e-71
mbrace2	3.730368e-01	2.113899e-02	1.764686e+01	1.075649e-69
cig_recY	4.814068e-01	2.977606e-02	1.616758e+01	8.539278e-59

Figure 4: Part of R output of logistic regression

Variables "combgest" and "dplural" are also significant here as using GUIDE, but some other variables appear here. It may not be safe to interpret the result, because interaction and multicollinearity cannot be simply ignored in linear models.

4.5 Xgboost

Xgboost has the same problem as logistic regression, so I use the same data as in logistic regression to avoid it. After tuning parameters, I get the lowest error rate on the validation set in Table 5.

Table 5: Error rate on the validation set using Xgboost

Booster	Error rate
CART tree	0.2679444
Linear	0.2762222

Xgboost is a very popular machine learning method because it is fast and accurate. However, it does not improve accuracy significantly here in this case, even worse than simple logistic regression. Furthermore, it uses a sparse matrix for computation, which recode each level of a categorical variable into 0 and 1, and that is why it is so fast. However, recoding step also turns it into a black box that we can hardly tell which variables are important.

5 Conclusion

It is hard to say which method is better as the differences in mean error rate are not statistically significant on the validation set. However, ensemble methods like GUIDE forest and xgboost using tree booster tend to have slightly lower error rates.

Another thing is that GUIDE bagging does not provide an expected lower error rate like GUIDE forest. It is reasonable because this method uses 100 pruned GUIDE classification trees, while forest uses 500. It is the same situation for logistic regression and xgboost. Therefore, I guess that more data results in lower error rates in this specific case.

6 Discussion

GUIDE classification tree is the most safely interpreted method, logistic regression is good for prediction, and xgboost is the fastest one. Ensemble methods are better than single methods in prediction. Logistic regression and xgboost will crash if the test sample contains class values that do not appear in the training sample.

Therefore, there is no silver bullet for choosing algorithms, as there is nothing better than a linear algorithm to catch a linear link, but there is nothing better than a linear algorithm to catch a linear link.

In this specific case, I recommend using GUIDE, because error rates differ a little, and I think it is really important to know what contribute to a low birth weight. Mothers can take precautions according to it to reduce the chance of giving birth to a low weight baby, which is practical for real-world applications.

Finally, the current analysis only uses a subset of the whole data set. It may be interesting to use more data for analysis.