Executive Summary - Qualitative Prediction of Birth Weight

Mohan Rajendran

2019-07-11

Overview

The purpose of this analysis is to perform qualitative prediction of birth weight based on various variables from mother's medical history like age, last_menstral_weight, race, smoking history, past preterm labours, hyper tension, uterine irritability, number of physician visits

Analysis Method

A BirthWt data consist of 159 rows. Owing to the small amount of data cross validation techniques are deployed to come up with a prediction model. A double cross validation is performed to select between various models.

Response Variable

The categorical variable 'low' is used as the response variable. Considering that the child birth weight depends on various external factors like diet, prenatal intake, herdity which are not covered in the data set, it is decided to do a qualitative analysis instead of quantitative. So the scope of the analysis is to estimate whether the child will be underweight or normal.

Audience/Practical Purpose

This model will enable healthcare professionals, doctors, midwifes and even the parents to predict the risk of a child being born underweight. Though the scope of model is not to predict exact birth weight, it still enables the user to categorize high risk pregnancies.

Methodology

Considering most of the variables are factors its decided to use Logistic Regression and Random Forst Classification. Variable selection is performed using step function which ignored age and number of physician visits. Also depending on the p value Uterine Irritability variable is removed as it is not significant. After performing double cross validation, Logistic Regression model is found to perform better with a classification rate of 0.7195.

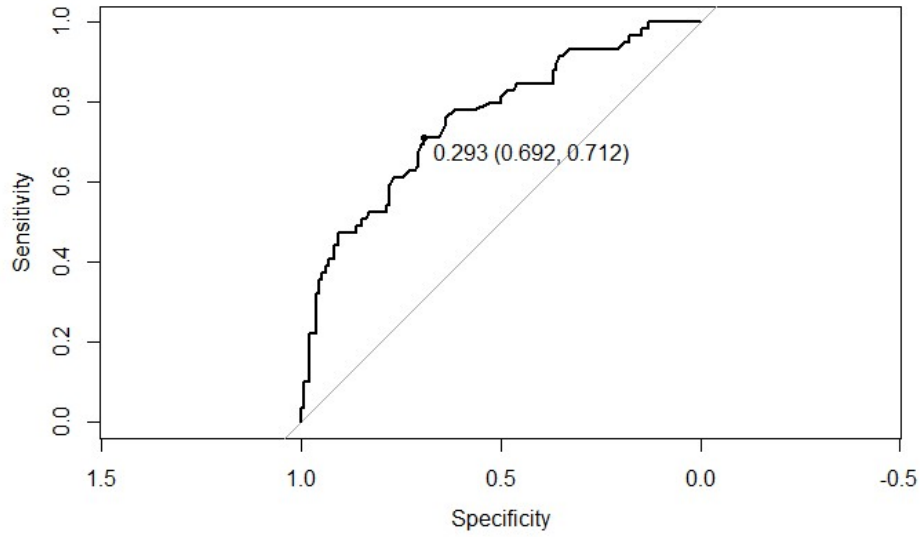
Confusion Matrix

```
          Actual
Prediction   0   1
         0 115  38
         1  15  21
```

Limitation

The model performs better when it comes to predicting normal weight child births than identifying low weights. Though the classification rate is 0.6772 this is most likely due to the fact normal weights (low = 0) forms the majority category of data. This is evident from the confusion matrix. Model performance can be improved by tweaking the threshold value to be less than 0.5.

ROC curve is computed by fitting the model to the full data and the best threshold is identified as 0.293.



Though this results in high number of false positives, it is better in this particular analysis as it gives the medical professionals to flag all potential cases without missing out on many true positives. The resulting confusion matrix is as follows

```
          Actual
Predicted  0  1
    FALSE 90 17
    TRUE  40 42
```