Logistic Regression in Spark:
Predicting low birth weight in children

Shane Stephenson
stephenson.shane.a@gmail.com

University of Maryland Global Campus
Summer 2019 DATA 650
Dr. Ozan Ozcan

Introduction

The purpose of the study on low birth weight is to take common variables in pregnancy and use them to predict whether a child will be born with low birth weight. Newborns who fall within this low birth weight category face several substantial obstacles after being born that parents and doctors need to be prepared for and ensure that the child gets the best shot at life. Low birthweight children frequently have issues with underdeveloped lungs and brains and sometimes need to be kept under hospitalization to stay alive and develop properly. Additionally, having a low birth weight can lead to a number of complications during delivery that require emergency medical treatment—such as breathing issues and an increased risk of broken bones during birth.

The independent variables that will be used to train the model are the mother's age, the mother's race, the mother's history of smoking while pregnant, the mother's history of preterm labor and hypertension, presence of uterine irritability in the mother, and the number of first trimester physician appointments by the mother. Race is treated as categorical variables. Smoking history, hypertension, and uterine irritability are binary variables and mother's age, history of preterm labor, and number of first trimester physician appointments are continuous numeric.

Data Exploration and Visualization

In the data exploration phase, the first step is to located missing values and deal with them appropriately. Fortunately, with this dataset, there are no missing values and there are only 189 rows of data. This dataset is also skewed, with about 2.2 times as many labels for normal birth weight than low birth weight.

Crosstabulation tables between birth weight category and all binary and continuous variables are generated and a very striking observation to be made is that 73% of African American mothers and 60% of other non-white mothers have children with low birth weight while only 32% of white mothers do. Also, notable (but not unexpected) is that a much higher proportion of mothers who smoke give birth to a child with low birth weight. Hypertension and uterine irritability are not applicable to most mothers, but they have a significant effect on birth weight when they are present.

Looking at the distribution plots for the continuous variables, a much higher proportion of low birth weight mothers have no first trimester physician appointments. There is also a significantly higher proportion of preterm labors in mothers who give birth to low weight children while age looks to be distributed about the same between the two birth weight categories.

Something to note about the continuous distributions is age is approximately normally distributed but both preterm labor history and first trimester physician appointments are skewed heavily to the left (preterm labor history particularly so). To help mitigate this, both skewed variables will be transformed with a square root transformation.

Simple Logistic Regression

A simple logistic regression model without any tuning is built and the results are shown in *Figure 1*. To build this model, the data is split into two stratified (by birth weight category) samples, one that contains about 80% of the dataset and the other which contains about 20% of the dataset.

```
Model evaluation for the train data:
Accuracy = 0.7414965986394558
Error = 0.25850340136054417
Precision = 0.6666666666666666
Recall = 0.34782608695652173
F1 Measure = 0.4571428571428571
Area under PR = 0.5513161786453712
Area under ROC = 0.6343090830822213

Model evaluation for the test data:
Accuracy = 0.6428571428571429
Error = 0.3571428571428571
Precision = 0.4
Recall = 0.3076923076923077
F1 Measure = 0.34782608695652173
Area under PR = 0.3686813186813187
Area under ROC = 0.5503978779840849
```

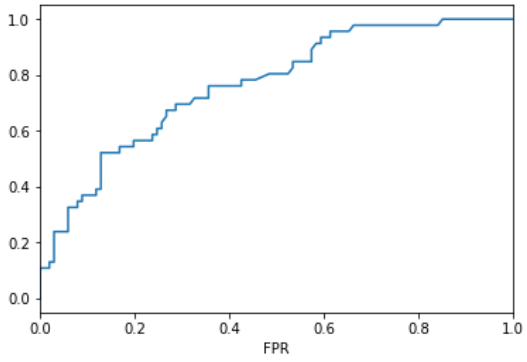

Figure 1 – Simple Logistic Regression Results

Looking at the results, the model is about 74% accurate on the training set and 64% accurate on the test set—so this model is definitely overfitting. However, since the data is imbalanced these are not the best metrics to judge the model on. Looking at precision, recall, and F1 score, it becomes obvious that this model is much better at predicting normal birth weight than low birth weight. This is expected because the data is imbalanced towards normal birth weight but also unfortunate because the goal of the model is to predict low birth weight—not predict normal birth weight. Looking at the ROC and the AUC, this model barely does better than random chance. AUC for a binary classification model that randomly guesses would be 0.5 and the AUC on the test set for this model is only 0.55.

Why Logistic Regression

For most binary classification problems (including this one), logistic regression is a good option. Logistic regression tends to be faster to train than most other classification methods—especially on larger datasets with a high number of features. Compared to other methods such as decision trees, support vector machines, and their various ensemble methods (random forests, gradient boosted decision trees, etc.), these large, high dimensional datasets are where logistic regression shines. Naïve-Bayes is the one of the other methods that can usually go toe-to-toe with logistic regression when it comes to speed on large datasets—with Naïve-Bayes frequently being faster.

Another big advantage logistic regression has is that it produces naturally well-calibrated probabilities. In fact, when trying to calibrate predictions from other models, logistic regression is commonly used on top of the predictions from other models to produce better probabilities. This method is known as Platt scaling.

Tuning the Logistic Regression Model

Using a weighted loss function allows our model to be tuned for increased performance. The natural way to weight a model so that it produces well-rounded results is to use a multiplicative factor of 1 when calculating loss for data in the majority class and use a multiplicative factor equal to the ratio of the majority class to the minority class when calculating loss for the data in the minority class. *Figure 2* shows the output results of using this weighting scheme to train a logistic regression model on the same training/test split as the simple model.

```
Model evaluation for the train data:
Accuracy = 0.6938775510204082
Error = 0.30612244897959184
Precision = 0.5081967213114754
Recall = 0.6739130434782609
F1 Measure = 0.5794392523364487

Model evaluation for the test data:
Accuracy = 0.5952380952380952
Error = 0.40476190476190477
Precision = 0.3888888888888889
Recall = 0.5384615384615384
F1 Measure = 0.45161290322580644
```

Figure 2 – Accuracy Metrics for Balanced Model

Comparing *Figure 1* to *Figure 2*, it is apparent that this tuning method works for making the model more balanced. While the overall accuracy of the model decreased, the F1 score and ability to predict the minority class increased. While this model is better than the first, it is still performing very poorly at capturing actual low birth weights. To get around this, the multiplicative factor described above will be taken as a variable instead of simply using the ratio of the majority class to the minority class. This variable will be iterated over with multiple models built and key metrics for each model captured. This will allow a weighting scheme to be found that captures the minority class much better while also not treating every row as if it were in the minority class. This method requires the data to be re-split into three sets: a training set, validation set, and test set. The training will be used to train each model, the validation set will be used for measuring the performance of each model, and the test set will be used to evaluate the final model selected.

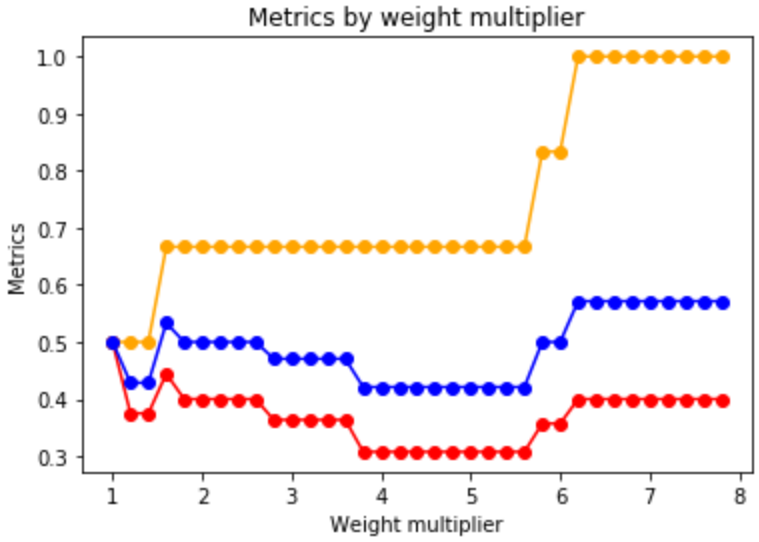

Figure 3 – Tuning the Logistic Regression Performance

Figure 3 shows the plot of the metrics for each model as a function of the weight multiplier. Yellow represents recall, blue represents F1 score, and red represents precision. Each dot represents a 0.2 increase from the previous weight. The graph at weight multiplier 6.2 shows a huge increase in recall but it is also close to the max value that precision takes as well. The graph also shows that 6.2 is the weight multiplier that produces approximately the max value for the overall F1 score in the plotted range as well (unsurprisingly, as F1 is the harmonic mean between precision and recall). Based on this, choosing a weight multiplier greater than or equal to 6.2 seems to be the best choice to suit the purpose of this model. The accuracy metrics for this final selected model are shown in *Figure 4*.

```
Model evaluation for the train data:
Accuracy = 0.5503875968992248
Error = 0.4496124031007752
Precision = 0.40425531914893614
Recall = 0.95
F1 Measure = 0.5671641791044776
Area under PR = 0.4019008741547089
Area under ROC = 0.6603932584269663

Model evaluation for the test data:
Accuracy = 0.4222222222222222
Error = 0.5777777777777777
Precision = 0.32432432432432434
Recall = 0.9230769230769231
F1 Measure = 0.48
Area under PR = 0.32296142296142294
Area under ROC = 0.5709134615384616
```

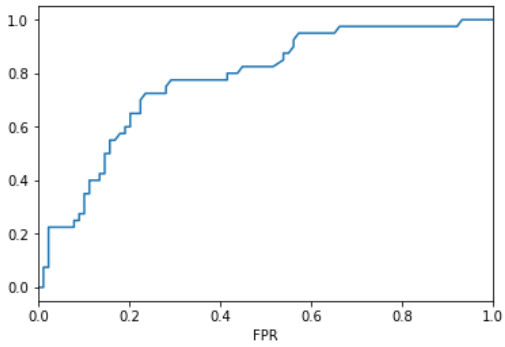

Figure 4 – Accuracy Metrics for Final Model

Looking at the output for the final model, the F1 score is clearly not the best seen so far on the test set. It is, however, the best recall seen so far. Only 32% of the children predicted to have low birth weight will have low birth weight—but almost all (about 91.6% on the test set) of the children born with a low birth weight will are predicted to have a low birth weight. When it comes to predicting whether a child will be born with a low birth weight or not, it is more important to predict low birth weight when there is low birth weight than it is to predict normal birth weight when the child has a normal birth weight. If the model predicts low birth weight and the child is born with a normal birth weight then it's really a "no harm, no foul" situation. It's better to prepare for the worst and hope for the best than to be underprepared when a child is born with a low birth weight. The model can be tuned even more precisely with more

information on the actual costs associated with misclassifying a normal birth weight as low birth weight.

Alternative ML Methods

Comparing the simple logistic regression model developed to the other simple models built, it is easy to see that without any model tuning other methods outperform logistic regression on the training and test sets. Tuning the logistic regression model produces a better model for our purposes, but similar methods can also be used to tune other models and produce comparable—or better—results. Particularly, with the small size and limited features of this dataset, training a decision tree using weighted binary cross-entropy loss should have good predictive power. Despite training ensemble tree models in the notebook, going with an ensemble tree model would not be recommended for this dataset because it is so small.

Similarly, the continuous numeric variables can be appropriately binned into categorical variables and a multinomial Naïve Bayes classifier trained on the data. Naïve Bayes classifiers have the distinct advantage that they can get relatively accurate predictions with small datasets—which would be perfect in this situation since this dataset is very small.

Generating synthetic data using a method such as SMOTE is another option to use with imbalanced data but with the limited range of most features and small size of this dataset it's unlikely that synthetic data would produce much of a performance boost compared to using a weighted loss function.