FETAL HEALTH CLASSIFICATION

Sidney Christensen

ISC 4242

Spring 2024

INTRODUCTION

Child mortality is an unwavering societal issue that is part of the United Nations' Sustainable Development Goals. It is expected that by 2030, preventable child and maternal mortality during labor will cease. As of 2017, maternal mortality is responsible for almost 300,000 deaths during and following the pregnancy. Over 90% of these preventable deaths took place in low-resource settings; thus, shedding light on the root of the problem at hand. With this issue in mind, it is proposed that Cardiotocograms (CTG's) may be implemented in low-resource settings to address and prevent child and maternal mortality where it is preventable. Cardiotocograms are simple and cost accessible tools for evaluating fetal health, a tool that has strong potential to break financial barriers in the healthcare industry for regions with less financial security/ resources. This problem aligns with my interests in public health, especially the labor and delivery sector. With this, exploring solutions where it is preventable fuels my strong interest in working with this fetal health dataset. The goal of this project to be able to build a model that can predict fetal health during the third trimester using cardiotocogram data as accurately as a highly trained doctor would.

DISCUSSION

This project's dataset is titled "Fetal Health Classification" and consists of 2126 observations (pregnant women all in their third trimester). The dataset is from the University of California Irvine Machine Learning Repository. Expert obstetricians have classified each observation into 3 classes: normal, suspect, and pathological. There are 21 predictors as follows: baseline value (Baseline of the fetal heart rate), accelerations (number of fetal heart rate accelerations per second), fetal movement (number of fetal movements per second), uterine contractions (number of contractions per second), light decelerations (number of LDs per

second), severe decelerations (number of SD's per second), prolonged decelerations (number of prolonged decelerations per second), abnormal short term (percentage of time with abnormal short term variability), mean value of short term (mean value of short term variability), and percentage of time abnormal (percentage of time with abnormal long term variability). The last ten predictors pertain to the Fetal heart Rate (FHR) histograms of each observation throughout the third trimester. The creators of this dataset made ten different measurements of the histogram to be predictors themselves. These predictors include histogram width, minimum, maximum, number of peaks, number of zeros, mode, mean, median, variance, and tendency. The data was recorded during the third trimester (the mother is hooked up to a monitor for some time before labor and during labor). The data is unique in that some predictors pertain to mom while others only pertain to fetus.

Seeing this is a classification problem, I will be exploring support vector machines, XGB, random forests, and neural networks. For the Support Vector Classifier, XGB, and Random Forest models, the ten most significant features determined through ANOVA are used. All features were still used for both Neural Networks as they performed better with all features included. There are only 3966 observations after applying SMOTE (Synthetic Minority Over-sampling Technique), which is not optimal for a Neural Network as they need much more data for adequate training. I used all features to also provide as much information as possible to the models. Thus, preprocessing the data will consist of SMOTE (all five models) and normalization (SVM and Neural Networks). SMOTE is necessary as the three classes are heavily imbalanced, 70% of the fetuses being "normal". The researchers using this dataset performed the Synthetic Minority Over-sampling Technique to avoid overfitting of the model on skewed classes during the training phase. The preprocessing pipelines for the Extreme Gradient Boosting and Random Forest models only consist of SMOTE as tree algorithms are not sensitive to feature scale. However, the Support Vector Machine and both Neural Networks will require feature scaling after SMOTE, as these models are sensitive to feature scale. In the code, I have provided a pipeline list for each model for clarity.

IMPLEMENTATION

This section gives a brief overview of the code. The data is first explored through a few visualizations, and then spit into testing (80%) and training (20%) sets. The ANOVA test is applied first on the training set, then SMOTE, then normalization for the models needing feature scaling. For the last two models, the Neural Networks, the data is loaded again in to use all features. For the first three models, feature selection is applied first because it assumes all samples are independent, an assumption SMOTE violates. Feature selection is only applied on the training set to prevent data leakage since feature selection will only be used in training. Normalization is last in the pipeline to ensure features are scaled for modeling. The code provides a clear pipeline for each model for further understanding of the preprocessing required for each model. A grid search for the best parameters for each model is performed for the first three models. Those parameters are then used to build a final model. For the Neural Networks, I wanted to take this opportunity to adjust and tweak the parameters myself to better understand the nature of neural networks. However, adjusting and building up both neural networks is a slow process as we do not want to overcomplicate the model. If a simpler model gives the same accuracy of a more complex model, the simpler model is desired as less is more. The implementation is described in complete detail in the code. Seeing that this is a matter of fetal health, high recall is valued more than high precision as false negatives are what we want to avoid most. Classification tables have been displayed for all five models to compare such evaluation metrics.

RESULTS

Using the classification reports from all five models, let us look at recall, accuracy, and F1-scores. While we want to be able to predict all classes correctly, it is crucial that the selected model can classify the “suspect” and “pathological” classes correctly as we previously mentioned that these two classes are the fetuses at risk, potentially needing medical intervention. All models performed well on the first class (“normal”) having recall and F1-scores all well above .90. We want to look at recall because for this domain, false negatives are costly. In terms of both recall and F1-scores, the leading models are the Random Forest and XGB. While the Random Forest has recalls of .95, .91, and 1.00, the XGB model has recalls of .96, .86, and 1.00 for classes 0 (“normal”), 1(“suspect”) and 2(“pathological”), respectively. Both the Random

Forest and XGB models had a recall of 1.00 for the "pathological" class, the Random Forest had a recall of .91 which is better than that of the XGB model (.86). Additionally, the models both had an accuracy of 95%. In terms of F1-scores, the Random Forest had almost similar scores to those of the XGB for the ("suspect") and ("pathological") classes. While the XGB model had a better precision for class 2 (.88), recall is more important to us. All this considered, the Random Forest model is slightly better in terms of recall, thus making it the final chosen model as it performs best for classifying fetal health.

CONCLUSION

As discussed in the "Results" section, it is decided that both the XGB and Random forests are promising models due to their strong recalls for all three classes and 95% accuracy. However, the Random Forest had a stronger recall for the "suspect" class than that of the XGB, making it the tie-breaking factor. Let me now summarize the knowledge gained from this project. By building five various predictive models for a cause such as this, there is much to takeaway. Domain knowledge is extremely helpful in knowing how to evaluate your models and which metrics should be considered most heavily. For example, fetal health classification as we know is a matter of life or death, so false negatives are detrimental in this circumstance. For that reason, I valued recall more than precision during the assessment of all five models. Next, in terms of data preprocessing, I learned which models require feature scaling and which do not. As seen in the code, the two tree-based models (XGB & Random Forest) do not require feature scaling, while the Support Vector Classifier and Neural Networks do. The order of steps in which the data is preprocessed is another crucial factor in the success and reliability of the models as our goal is to prevent data leakage from beginning to end. For example, feature selection through ANOVA was only performed on the training data as so that information from the test set does not interfere with training. Lastly, this project allowed me to build a strong foundation in Neural Networks by forcing me to learn which parameters work best together, which parameters are most appropriate for your specific problem, and the order in which we should be adjusting those parameters. To be specific, an example of this is learning that using the "soft max" activation function works best with the Cross Entropy Loss function since it is built to work with those probabilities unlike hinge loss. Additionally, I learned how different optimization algorithms work and why we

should use them. Seeing the ways in which two different architectures perform allows me to see the degree to which neural networks are nuanced and their sensitive nature. With neural networks, less is more, but accuracy is most desired, and sometimes the best accuracy is achieved with a more complex architecture.

Overall, we see that predictive modeling/machine learning is extremely nuanced in nature, allowing for us to educate ourselves and research the most optimal methodologies that work for your specific problem at hand. There are so many features to consider and address. Is the problem a matter of classification or regression? If classification, are the classes balanced? If we want to perform feature selection, are our variables numerical, categorical, or both? This determines which method of feature selection best suits our data. What is the domain of our problem and what outcomes matter most? Asking and addressing these questions make all the difference as we tailor our approach for every project for the best outcomes. Data is never conventional, and it is our job as data scientists to be so careful with it through every step of the data life cycle, with the end goal being to make a difference in our society with reliable insights and successful models.

REFERENCES

Fetal Health Classification (kaggle.com)

Use of Machine Learning Algorithms for Prediction of Fetal Risk using Cardiotocographic Data - PMC (nih.gov)