Predictive Modeling for Low Birth Weight Classification

Brandi Jones

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

Faculty Advisor: Dr. MinJae Woo

INTRODUCTION

- Infant birth weight is known to be a good predictor of clinical outcome.
  - *Birth weights less than 2500 g are classified as low birth weight (LBW).*
  - **Low birth weight has been linked with increased infant morbidity and mortality risk, with the smallest infants most at risk.**
- Prediction of low birth weight serves as a valuable preventative tool.
  - **Identification of at-risk infants is key for early and effective clinical intervention.**
- Many factors have been linked to LBW including preterm birth, maternal and paternal health and lifestyle factors, maternal age, and access to prenatal care.
- **The purpose of this study is to explore the use of current modeling methods for infant low birth weight prediction using a variety of maternal and paternal factors.**

METHODS

Dataset

- **Infant dataset was obtained from the National Survey of Family Growth (NSFG) survey** conducted by the Centers for Disease Control and Prevention (CDC) from 1973-2019.
  - Survey collects information on fertility, family planning, and reproductive health in the United States.
  - The sample was designed to be representative of live births in the United States using continuous interviewing/fieldwork survey methodology.
  - **Dataset included 101,400 live births and 41 variables.**

Data Processing

- **Low Birth Weight (LBW) binary classification response variable created using 5.511557 lbs (2500 g) as threshold.**
- MICE Imputation performed for missing values after handling of coded missing.
- Use of 60:20:20 ratio for train/validate/test sets for all models.

Modeling Methods Used

- XGBoost
  - Hyperparameter Tuning
  - **Code adapted from Dr. MinJae Woo DS7140 Notes**
- Naïve Bayes
- Random Forest
- Logistic Regression

Modeling Results

- AUC/ROC Curves calculated for each model.
- Confusion Matrix created for each model.
- Accuracy, F1 Score, Precision and other model performance metrics calculated.

Brandi Jones
MSDSA Graduation
Spring 2025

References

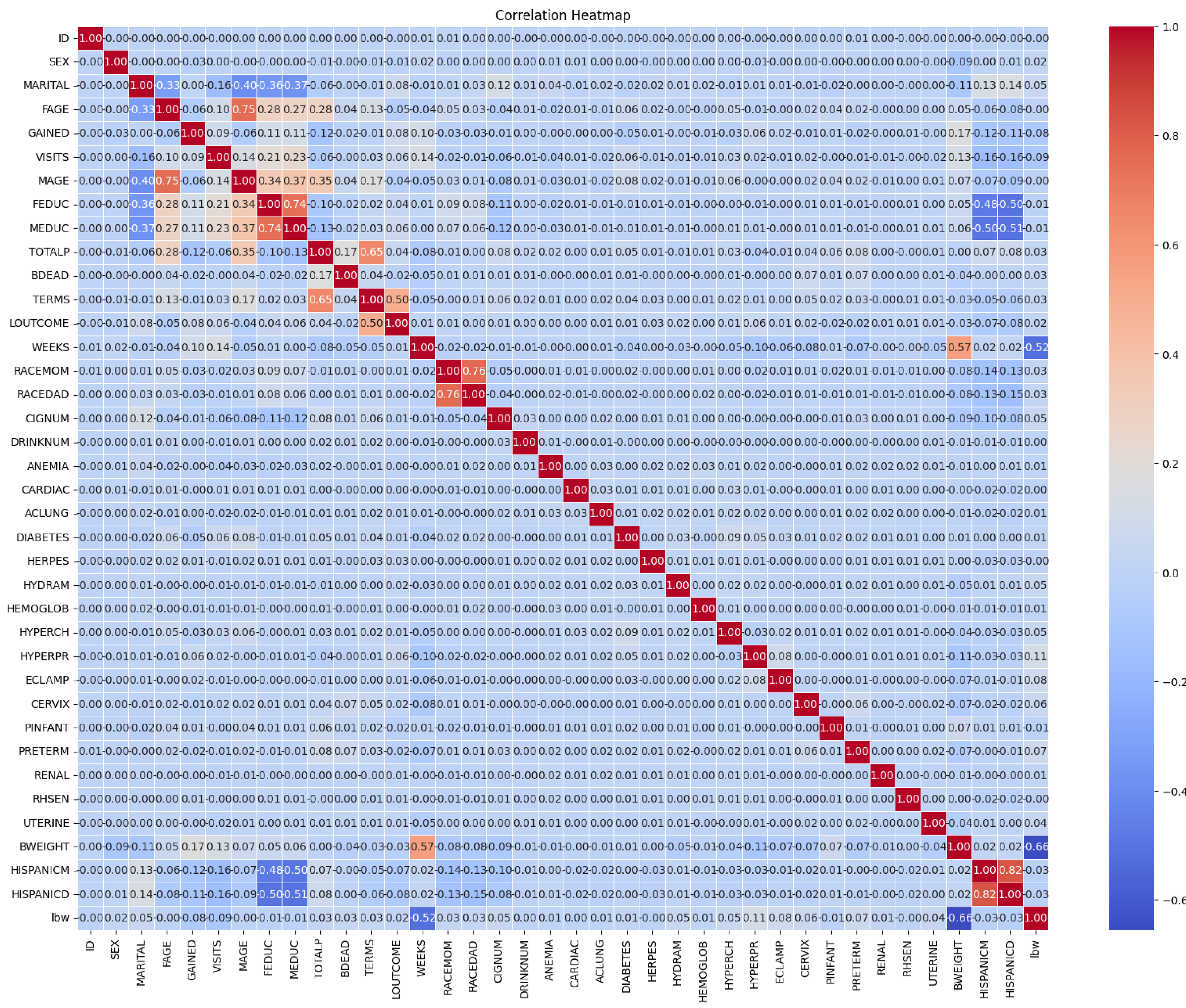

Figure 1: Correlation Heatmap of Explanatory Variables

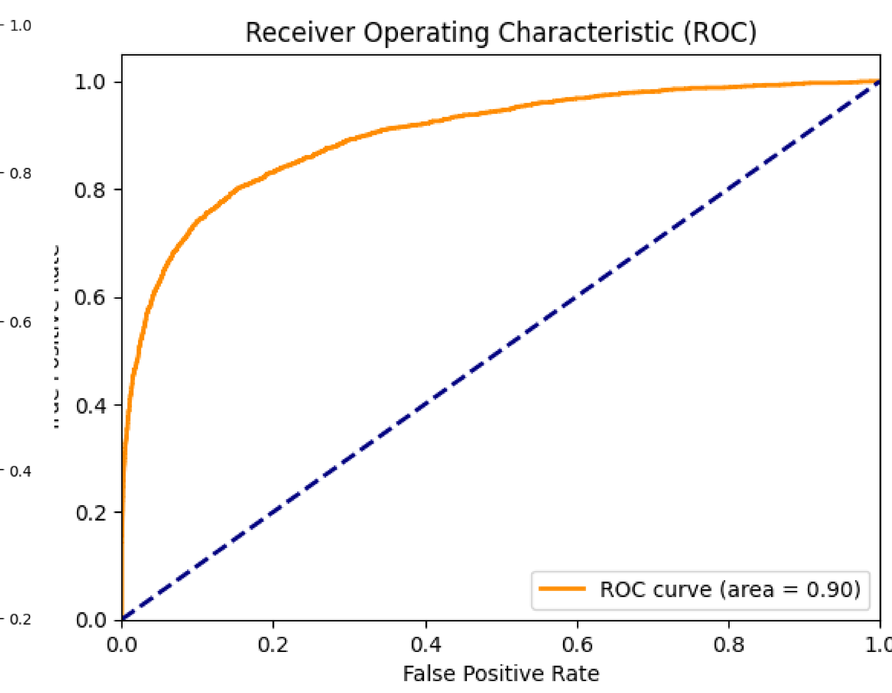

Figure 2a: ROC Curve from XGBoost Classification Model

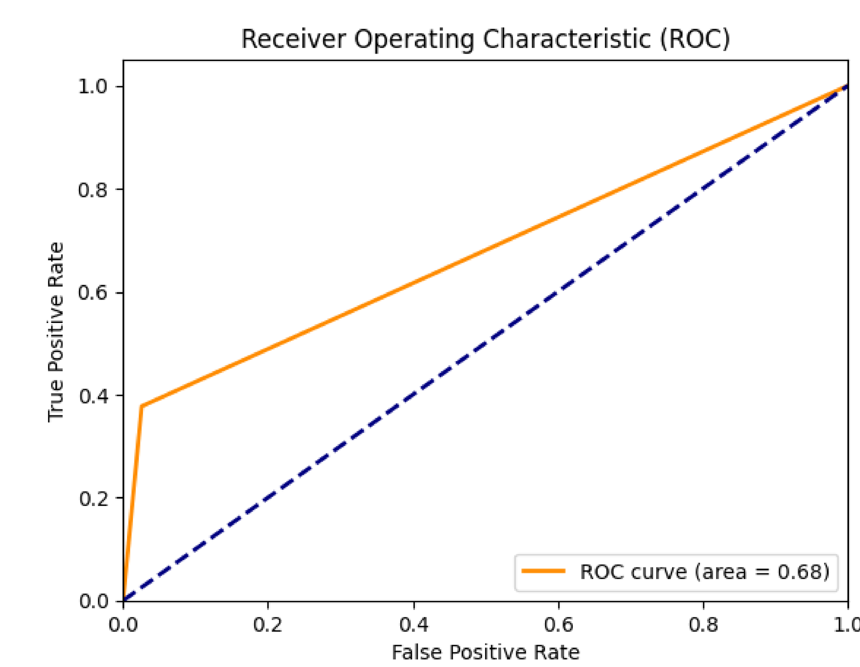

Figure 2b: ROC Curve from Naïve Bayes Classification Model

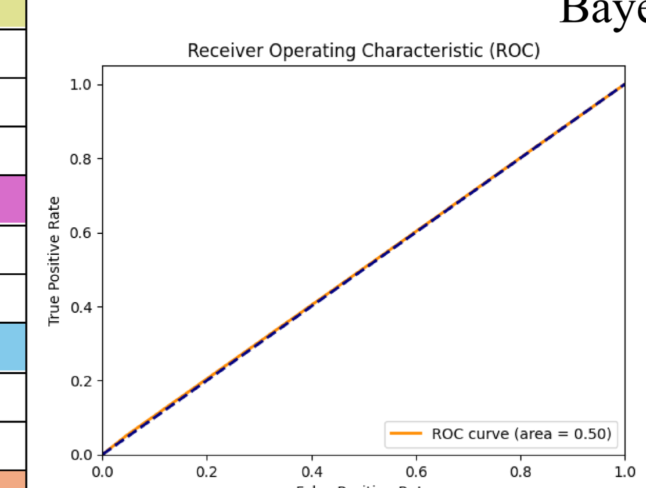

Figure 2c: ROC Curve from Random Forest Classification Model

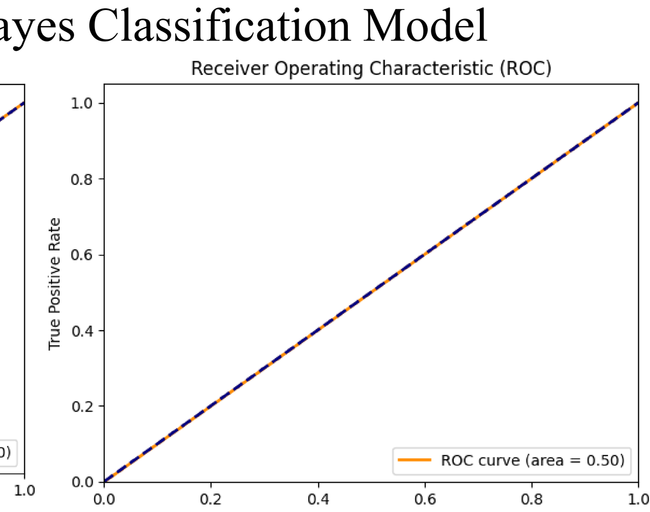

Figure 2d: ROC Curve from Logistic Regression Classification Model

Table 1: Confusion Matrix Output for Each Model

		Actual	
		Naïve Bayes Confusion Matrix	
		Positive	**Negative**
Predicted	**P:**	18157	479
	N:	1024	620
		Random Forest Confusion Matrix	
	P:	17784	852
	N:	1560	84
		Logistic Regression Confusion Matrix	
	P:	18604	32
	N:	1643	1
		XGBoost Confusion Matrix	
	P:	17792	844
	N:	1566	78

Table 2: Model Performance Metrics Comparison

	AUC	ACCURACY	F1 SCORE	PRECISION	SENSITIVITY
XGBOOST	0.8994	88.12 %	0.060795	0.084598	0.04745
NAÏVE BAYES	0.6757	92.59 %	0.4520598	0.564149	0.37713
RANDOM FOREST	0.5027	88.11 %	0.0651163	0.089743	0.05110
LOGISTIC REGRESSION	0.4994	91.74 %	0.0011923	0.030303	0.00061

Figure 3: Final Model Feature Importance

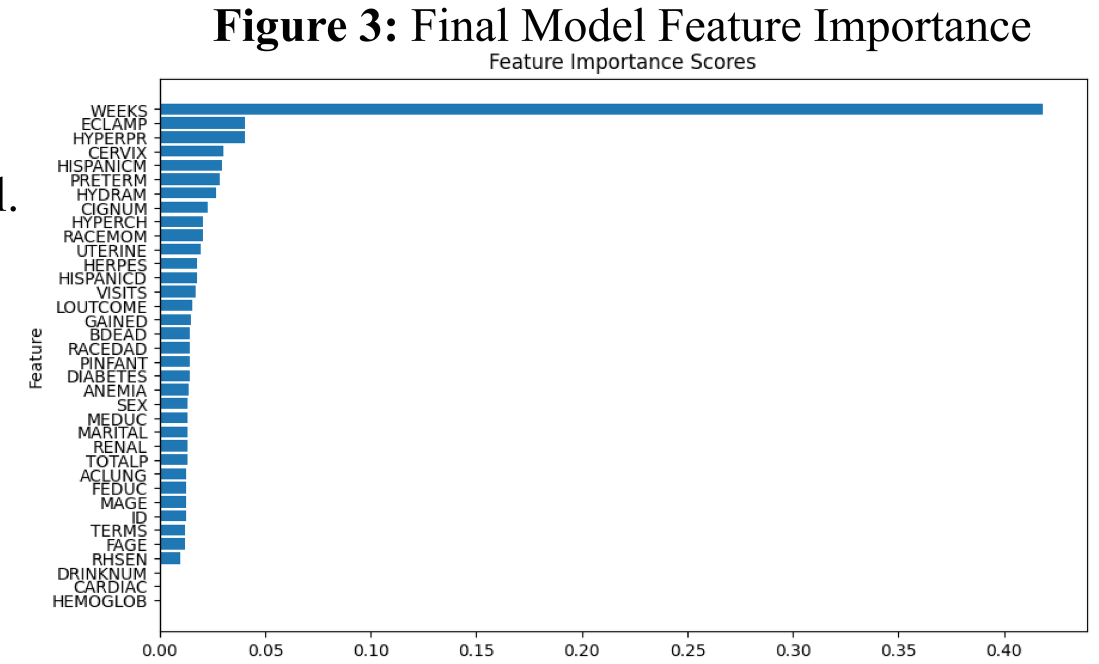

Table 3: Final XGBoost Hyperparameters

	XGBoost Best Hyperparameter Values
0	colsample_bytree: 0.825217
1	gamma: 0.002147
2	learning_rate: 0.015088
3	max_depth: 8.000000
4	min_child_weight: 5.000000
5	n_estimators: 399.000000
6	subsample: 0.679022

Table 4: Final Model Feature Importance

Top 5 Feature Importance Contributions for XGBoost Model	
Completed Weeks Gestation	Importance Score: 0.41825124621391296
Mother has/had Eclampsia	Importance Score: 0.04032757505774498
Mother has/had Preg. Hypertension	Importance Score: 0.040144454687833786
Mother has/had Incompetent Cervix	Importance Score: 0.029999906197190285
Mother is Hispanic	Importance Score: 0.029487695544958115

RESULTS

Model Outcomes (Table 2)

- *XGBoost had the highest AUC/ROC curve score of all models.*
  - Logistic and Random Forest models AUC/ROC indicate prediction is not better than random selection.
- *Naïve Bayes demonstrated the highest accuracy percentage.*
- F1 Score significantly higher in Naïve Bayes compared to others.
- Precision and Sensitivity were highest in Naïve Bayes (by large margin).

XGBoost Best Hyperparameters

- Tuning and selection of Hyperparameters calculated from max 30 combinations. Best parameters listed in **Table 3.**

XGBoost Feature Importance

- *Number of weeks gestation completed had an importance score of approximately 0.4183, indicating it contributes about 41.83% to the model's predictions.*
  - The next largest importance score was for "mother has/had eclampsia" (0.04033). **Table 4.**

DISCUSSION

- **Based on XGBoost feature importance scores, maternal factors contribute more to XGBoost model predictions than paternal factors.**
  - Completed weeks of gestation largest contribution.
    - Unsurprising due to current literature knowledge.
- **Model comparisons indicate overall XGBoost is the best model for predictive performance and discrimination between classes.**
  - *Naïve Bayes model best if focus is on accuracy and balance between precision and recall (F1 Score).*
    - Accuracy should be used with caution due to class-imbalanced dataset; not indicative of predictive ability.
  - ***XGBoost more adept at distinguishing between infants with low birth weight and those without, while Naïve Bayes achieves a higher proportion of correct predictions overall and more success in correctly identifying infants with LBW.***

Clinical Implications

- **XGBoost and Naïve Bayes** are the superior models compared to Random Forest and Logistic Regression, but **choice** between the two is **dependent on interpretation and clinical setting needs.**
- Ability to discriminate between classes advantageous in scenarios where identifying high-risk infants is needed for targeted interventions.
- **Feature Importance output is the most actionable finding from the study and was uniform across the models (Figure 3).**
- XGBoost predictions could help prioritize resources for prenatal care or implement preventive measures for mothers at higher risk of delivering low birth weight infants.
- **Naïve Bayes' higher accuracy and precision preferable in development of screening programs aimed at confidently identifying LBW infants.**
  - Relative model simplicity makes it ideal with limited computational resources.

Limitations

- Regardless of model performance, ability to interpret is crucial for clinicians' acceptance.
- Predictive modeling in healthcare warrants ethical considerations as regards biases in the data or algorithms.