Predict Infant Health Using the US Birth Data

Jinyuan Qi Wanru Xiong
Office of Population Research, Princeton University

PRINCETON UNIVERSITY

Research Question

Infant health is crucial for the health, growth, and development of people to their full potential in later life. Understanding the risk factors of infant health has been a long-standing goal in medical science and population research.

- What are the important predictors of infant health? What are the risk factors?
- What is the best prediction model?
- How could machine learning methods help us to select features and discover latent patterns in the birth data?

Data

The U.S. Birth Data 2017 (NCHS / CDC)

- National de-identified individual-level birth data
- Size: 3,864,754 observations

Features

- Demographic characteristics of the parents:
  - age, race, marital status, education attainment, birth order, and birth interval
- Medical and public services utilization:
  - prenatal care, time and place of birth, method of delivery, and payment source for the delivery
- Maternal lifestyle and health characteristics:
  - mother's height, weight, smoking behaviors and other risk factors during the pregnancy

Outcomes

- 5-minute Apgar score: score 0 - 10
- Birth weight: Normal: > 2500, Low: <2500 & > 1500, Very low: < 1500
- Abnormal conditions of the newborn: any of six conditions
- Congenital anomalies of the newborn: any of twelve conditions
- Infant living at the time of report: binary
- Infant breastfed at discharge: binary

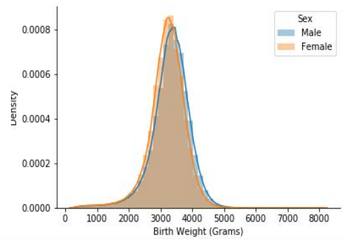
Preprocessing and Feature Engineering

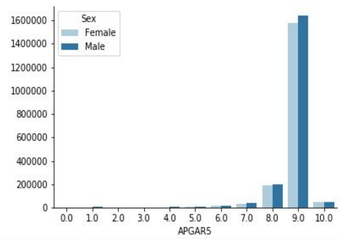
- Missing values
- One hot coding for categorical variables
- Nonlinear effect: squared term
- Combine sparse classes

Methods

- Unsupervised learning:
  - PCA
  - K-means
- Prediction - Classification:
  - Gaussian Naive Bayes
  - Multinomial Naive Bayes
  - Logistic Regression
  - Random Forest
  - Gradient Boosting
- Prediction - Classification:
  - OLS
  - Ridge Regression
  - Lasso Regression
  - ElasticNet Regression
  - Random Forest

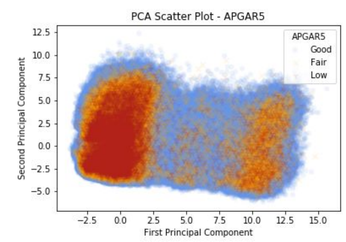
Distribution and Regression

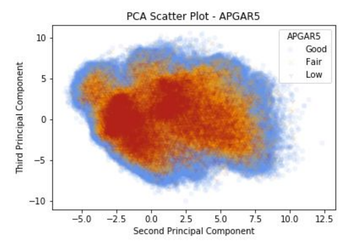




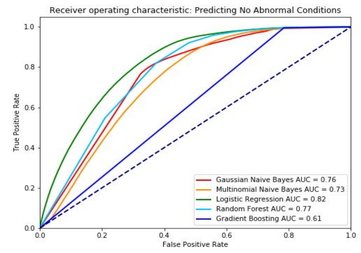
Outcome	Birth weight				5 min APGAR score			
Models	Train MSE	Test MSE	Train R2	Test R2	Train MSE	Test MSE	Train R2	Test R2
OLS	**157082**	**157149**	**0.55**	**0.55**	**0.54**	**0.53**	**0.20**	**0.20**
Ridge	**157081**	**157149**	**0.55**	**0.55**	**0.54**	**0.53**	**0.20**	**0.20**
Lasso	159585	159638	0.55	0.54	0.67	0.66	0.00	0.00
ElasticNet	159585	244688	0.55	0.30	0.67	0.66	0.00	0.00
Random Forest	203378	203401	0.42	0.42	0.57	0.56	0.15	0.15

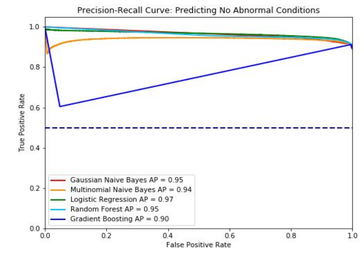
PCA



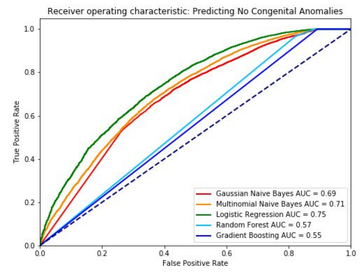


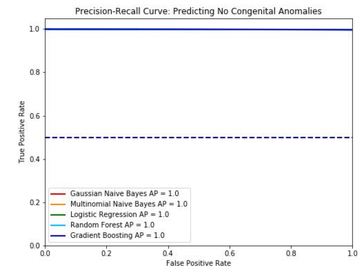
No abnormal conditions



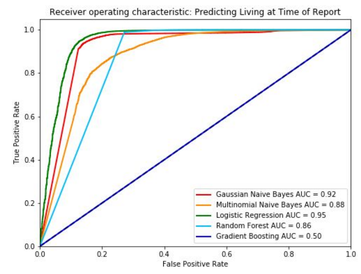


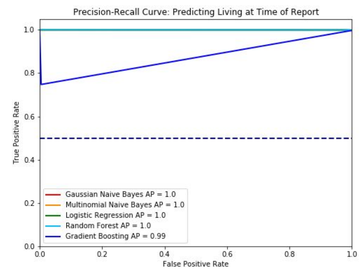
No congenital conditions





Infant living at the report





Infant breastfed at discharge