# Predicting Infant Health Using the US Birth Data

**Jinyuan Qi**
Office of Population Research
jinyuanq@princeton.edu

**Wanru Xiong**
Office of Population Research
wanrux@princeton.edu

## Abstract

In this project, we used machine learning methods to study infant health in the US based on the national birth data with more than 3.8 million newborns in 2017. Our goal is to find good prediction models and the determinants for infant health, measured by six key outcomes - infant living at the time of report, 5-minute Apgar score, birth weight, no congenital anomalies, no abnormal conditions, and breastfed at discharge. We used supervised learning models to identify important predictors and risk factors, and unsupervised learning to discover the latent profile of pregnancy. We found that Logit model performs the best in predicting binary outcomes and the Ridge regression works well for continuous outcomes. We identified several risk factors of infant health that concerns sociodemographic features of the mother, prenatal care, and methods of delivery.

## 1 Introduction

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. [9] [3] The National Center for Health Statistics (NCHS) and the Centers for Disease Control and Prevention (CDC) [6] provide national de-identified individual-level birth data, which recorded every birth in the US since 2005. The birth data are extracted from birth certificates that contain relevant information about demographic features and prenatal health. The size (total population of interest) and quality (consistent as required by law) of the data provide an excellent opportunity for us to study infant health at the population level.

In this project, we used machine learning methods to approach these classical questions about infant health: what are the most important predictors of infant health? What are the risk factors? What is the best prediction model for the birth data? How could machine learning methods help us to select features and discover latent patterns in the birth data? The answers to these questions would help us to better predict the risk of poor infant health at earlier stage and to identify the target group for effective policy interventions in advance.

To answer these questions, we used both supervised and unsupervised learning methods. We considered and compared four classification models (Gaussian Naive Bayes (GNB), Multinomial Naive Bayes (MNB), Logistic Regression (Logit), snd Random Forest (RFC) for four binary outcomes (living at the report, no abnormal conditions, no congenital anomalies, breastfed at discharge) and five regression models (Ordinary Least Squares Regression (OLS), Ridge Regression (Ridge), Lasso Regression (Lasso), ElasticNet Regression (ENR) and Random Forest(RFR)) for two continuous outcomes (Apgar score, birth weight). We carried out Principal Component Analysis (PCA) and Factor Analysis (FA) to do dimension reduction, and K-means clustering based on the reduced factors to discover latent profiles.

The results show that Logit model performs the best in predicting binary outcomes and Ridge regression works well for continuous outcomes. Certain measures are strong predictors of infant health, such as mother's weight gain during pregnancy, gestation period less than 27 weeks. We find some sociodemographic features that are associated with poor infant health outcomes, such as mother being Black and few prenatal care visits. The method of delivery also matters, in that Cesarean section is associated with higher risk of infant death.

The rest of the article is organized as follows: Section 2 reviews related research on infant health and other applications of the US birth data. Section 3 provides a description of the data source, features, and summary statistics of the key outcomes. Section 4 states all the models and evaluation metrics. Section 5 illustrates main results and section 6 ends with conclusions and discussions.

## 2 Related Work

Traditional studies on infant health largely relied on clinical trial or survey data. Such studies usually aimed to identify a causal link between one risk factor and a specific outcome of infant health, for example the effect of smoking on birth weight. Classical study designs include randomized clinical trial, case-control study, cohort study, and other identification strategies such as regression discontinuity [5] and difference in difference design [4]. These studies provide convincing evidence of various risk factors and social differentials of infant health. This project tries to incorporate these knowledge to build a prediction model for infant health using population data. Compared to survey data, our data have larger sample size yet less features. In fact, the data have entire population of interest instead of a sample. The data include many important risk factors of infant health found in previous studies, for example multiple birth, maternal age, and maternal BMI.

Many studies used the U.S. birth data to explore US infant health. Callaghan et al.(2000) used the data to study the recent decrease in US infant mortality rate between 2007 to 2013, and attributed the improvement to changes in the distribution of gestational age at birth [3]. Almond and Edlund (2007) used the data to test the Trivers-Willard hypothesis about socioeconomic status and sex ratio at birth and found married, better educated and younger mothers bore more sons and infant deaths were more male if the mother was unmarried and young. [1]. The recent birth data had also been used to illustrate the trends in differences in infant mortality between black and white [11]. These studies focused on one or two predictors of infant health bud do not show relative importance of different preditors in determining the infant health.

The application of machine learning methods in the study of infant health is growing. Chen et al. (2011) used neural network and decision tree to identify the risk factors of preterm birth among thousands of features from a survey dataset of 910 mother-child pairs in Taiwan, and found multiple birth is the most important risk factor for preterm birth [2].

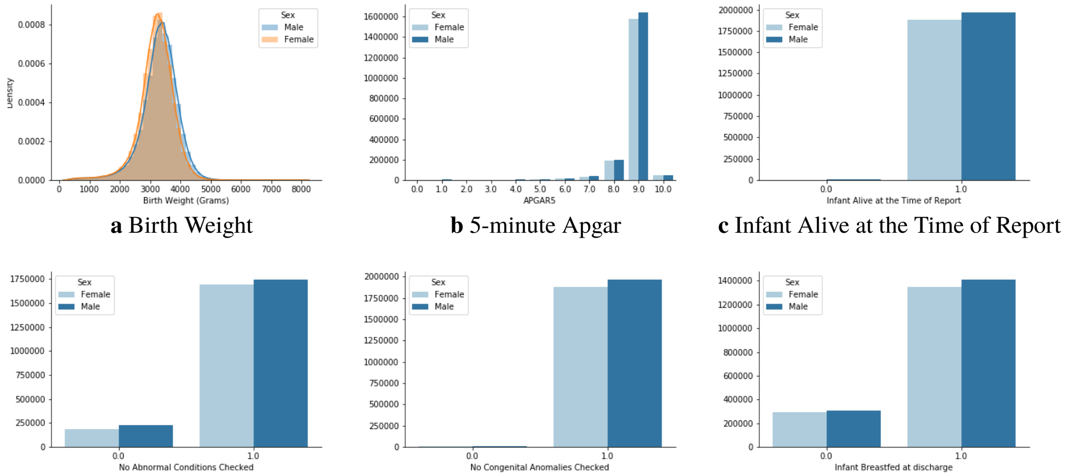
Using machine learning methods on US birth data, we can examine the previous findings using population data, rank the importance of features, and build prediction models based on a large training set.

## 3 Data

### 3.1 Features

The US birth data we use for this project are collected by NCHS and CDC [6]. Since US state laws require birth certificates to be completed for all births, and Federal law mandates national collection and publication of births data, the de-identified individual-level data are publicly available for research. We focus on the most recent birth data in 2017, which contain 3,864,754 observations. The features are abstracted from standard birth certificates filed in vital statistics offices of each state. The 240 features cover demographic and health variables of the mother as well as the birth. Specifically, the features include:

- 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;



**a** Birth Weight **b** 5-minute Apgar **c** Infant Alive at the Time of Report

**d** NO Abnormal Conditions **e** NO Congenital Abnormalities **f** Infant Breastfed at Discharge

Figure 1: **Distribution of the main outcomes by infant sex**

- Infant health characteristics: period of gestation, birth weight, Apgar score and records of abnormal conditions.

### 3.2 Key outcomes

Our six key outcomes of interest are 5-minute Apgar score, birth weight, abnormal conditions and congenital anomalies of the newborn, infant living at the time of report, and infant breastfed at discharge. The first two outcomes are continuous whereas the rest are binary. Figure 1 shows the distribution of the outcomes by sex of the new born.

The 5-minutes Apgar score measures "the need for resuscitation and a predictor of the infant's chances of surviving the first year of life" [8]. It is a summary measure of the infant's condition based on heart rate, respiratory effort, muscle tone, reflex irritability, and color 5 minutes after delivery. A score of 7 or greater indicates good to excellent physical condition. The distribution of Apgar5 is highly right skewed.

Birth weight is an important indicator of infant health. According to ICD-9 and ICD-10, birth weight less than 2500g is defined as low, and less than 1500g as very low. The distribution of birth weight is approximately normal.

Abnormal conditions record if the newborn has any one of the six serious conditions such as assisted ventilation and NICU admission. Congenital anomalies record if the newborn has any one of the twelve anomalies such as Down syndrome and cyanotic congenital heart disease.

## 4 Methods

### 4.1 Preprocessing and Feature Engineering

The quality of the US birth data is ensured by the standard data collection procedure and the uniform registration of the birth events. The number of features (241) is quite small ($n \gg p$) compared to the number of observation (around 3.8 million). In the data, around 11 % of the values are missing. We found that observations with more missing values are more likely to have poor health outcomes, therefore our data preprocessing primarily focused on coding and imputing missing values. We adopted different protocols for predictors and outcomes. For missing values in the predictors due to non-reporting and unknown, we generated a separate category when the variable is categorical, whereas we coded them as "not reporting" or the mode of the values when a variable is about reporting a rare condition. When the outcome variables have missing values, we deleted the observations for the specific analysis of that outcome.

We converted all the categorical variables to binary variables using one hot encoding. We also added squared terms for features that might have nonlinear effects on the outcome, such as the number of prenatal care visits. We combined highly correlated and overlapping variables such as racial categories and Hispanic origins. We dropped the features with more than 50 % missing values and deleted features with repeated information but a very detailed subcategories (e.g., the race variable with 31 categories to include every combination of mixed races such as "Black, AIAN, Asian, NHOPI, and White"), because the binary variables of those rare features would have very low variance.

### 4.2 Classification

We first split the data into training(70%) and test set(30%). We fitted four classifiers on the training set, and then use the models to predict the outcomes on the test set. The training time is a major concern in our analysis since our data size (n) is huge. We tried some advanced classifiers (e.g., Support Vector Machine, Neural Networks, etc.) and encountered dead kernel problem multiple times after running the functions for more than an hour. This suggests that computational cost/time is a major challenge in the application of those classifiers on big data. Further, the performance of neural networks is worse than logistic regression with less interpretability, thus we only reported performance of four classifiers with reasonable training time from the SciKitLearn Python libraries [10]. All parameters used are the default unless specified.

- **GNB:** A naive Bayes model assumes that the features are conditionally independent given the class label. Gaussian distribution is assumed for each class in GNB.
- **MNB:** Also a naive Bayes model but assumes that the class distribution is multinomial.
- **Logit:** The conditional distribution is a Bernoulli distribution rather than a Gaussian distribution in linear regression, because the dependent variable is binary. Logistic regression also assumes that the input observations are independent of each other. Both 'L1' and 'L2' penalty were fitted.
- **RFC:** RFC tries to de-correlate the base learners by learning trees based on a randomly chosen subset of input variables as well as a randomly chosen subset of data through bootstrapping [7]. To tune the hyper-parameters of RFC, we randomly selected a 10% subsample of our data and used grid search (`GridSearchCV` from SciKitLearn library [10]) to find the best combinations of the hyper-parameters. We then used the best set of the hyper-parameters on the entire training data set. More details are available in the Results section.

We used the following metrics to evaluate and compare performance of the models: training and testing accuracy, Area under ROC curve(AUC), Precision-Recall curve, macro average precision, recall and F1 scores.

### 4.3 Regression

We again split the data into training (70%) and test set (30%) and fitted the following regression models (from the SciKitLearn Python libraries [10]) on two continuous outcomes - 5-min Apgar score and birth weight (kg). All parameters used are the default unless specified.

- **OLS:** minimizes this loss function $L_{OLS}(\hat{\beta}) = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2$
- **Ridge:** minimizes $L_{Ridge}(\hat{\beta}) = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_2^2$ where $\lambda$ parameter is the regularization penalty. We tuned regularization strength (`alpha` in the function).
- **Lasso:** The loss function is $L_{Lasso}(\hat{\beta}) = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1$ We tuned regularization strength (`alpha`) and the maximum number of iterations (`max_iter` in the function).
- **ENR:** combines Lasso (L1) and Ridge (L2) regularization with the following loss function. We used grid search to tune `alpha`, `max_iter` and `l1_ratio` which is the ratio of L1 penalty in the combination. $L_{ENR}(\hat{\beta}) = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 + \lambda_2\|\hat{\beta}\|^2 + \lambda_1\|\hat{\beta}\|_1$.

- **RFR:** Similar to RFC, we used grid search to tune hyper-parameters of RFC in 10% sample. However, we could not implement the best combination of the hyper-parameters from the grid search because of high computational costs (e.g, running time is more than three hours when `max_depth=30`, `n_estimators = 200`). Thus, we presented results of RFR with sub-optimal hyper-parameters.

We used the training time, Mean Squared Error (MSE) and R-squared in the train and test set, and AIC/BIC for model comparison.

### 4.4 Unsupervised learning

We conducted PCA and FA to do dimension reduction and K-means clustering based on both the full set of features and the reduced components to discover the latent types of pregnancy. We tried K-means clustering after PCA because the huge data size and large number of features makes the direct K-means clustering very computational costly. PCA transforms a set of observations of possibly correlated variables into a set of values of linearly uncorrelated principal components. K-means clustering then partitions the observations in the lower dimensional space. Similar approach applies to FA.

To evaluate the results of K-means clustering after PCA and FA, we split the data to training set and test set to see if the latent patterns we found in the training set makes sense in the test set. We use reconstruction errors and log-likelihood on both training and test set to evaluate the models. We also compare the results of K-means clustering based on all features with that based on the principal components to evaluate the performance of PCA.

## 5 Results

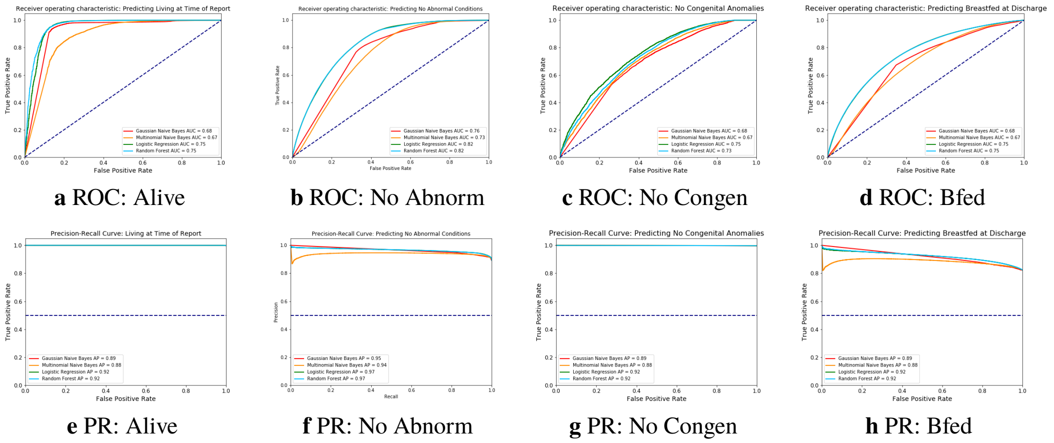
### 5.1 Classification

Figure 2 presents the ROC curves and the PR curves of all four models and Table 1 reports the evaluation metrics of the classification results. The model with the best performance (high training and testing accuracy, large AUC and high F1 score) is Logit across four binary outcomes. Logit with L2 penalty has consistent better performance than L1 penalty, and uses significantly less training time, thus we only presented results with L2 penalty in the Table 1. RFC's performance is very much similar to Logit in terms of accuracy and AUC, but often has lower recall and F1 scores. Further, RFC's training time was considerably longer than Logit in some models (e.g., no abnormal conditions) and required careful tuning. The hyperparemeter optimization of RFC is presented in Table 1. The training time scales up with increasing depth, smaller minimum sample splits and greater number of estimators.

We presented the feature importance of four binary outcome variables in Appendix Figure 4. We also used `statsmodels` package in Python to calculate the confidence intervals of the coefficients of two Logit models (living at the report and no abnormal conditions), which are partially presented in Table 1. The function failed to compute confidence intervals for other Logit models because of high computational costs.

To predict infant mortality/survival at the time of report, weight gain of the mothers during pregnancy and enough gestation period beyond 27 weeks, and no use of cesarean section (C-section) are the most important features. In the logistic regression, we also found that pre-pregnancy diabetes, excessive prenatal health-care visits are statistically significantly associated with higher risks of infant mortality, while joining the WIC food program (which is a special supplemental nutrition program for low-income women, infants, and children) and moderate number of prenatal care visits are associated with lower risks of infant mortality.

For the prediction of no abnormal conditions at birth, gestation period over 37 weeks, and no use of steroids and chorioamnionitis in labor and delivery are the most important predictors. Both of the Logit models show that male infants have higher risks of mortality and having abnormal conditions than female infants. In addition, infants born in the hospital (instead of home or other places), mothers with greater height, and having enough prenatal-care visits are significantly associated with lower risks of having abnormal conditions. Race also plays a role in the predictions, as the Table 1



**a** ROC: Alive

**b** ROC: No Abnorm

**c** ROC: No Congen

**d** ROC: Bfed

**e** PR: Alive

**f** PR: No Abnorm

**g** PR: No Congen

**h** PR: Bfed

Figure 2: **Receiver Operating Characteristics (ROC) curves and Precision-Recall (PR) curves of four outcomes using five classifiers** - ALIVE means infant living at the report; NO ABNORM means no abnormal conditions; NO CONGEN means no congenital conditions; BFED means infant breastfed at discharge. Colors of the curves: GNB, MNB, Logit, RF.

shows that black infants, regardless of their ethnicity (whether Hispanic or not), have higher risks of having abnormal conditions, while American Indian or Alaska Native, Asian, and white Hispanics have lower risks than non-Hispanic whites. Mother's pre-pregnancy diabetes, infections with syphilis and and Hepatitis C have large negative coefficients and small p-values, which implies statistical significance in predicting abnormal conditions. As for congenital anomalies, chromosomal disorders (e.g., Down's syndrome) reported are the most important predictors.

The feature importance of breastfed at discharge has less skewed distribution with more sociodemographic variables as the most important features. For instance, mothers who were married non-smokers, using medicaid or private insurance for payments are more likely to breastfeed their babies at discharge. Some variables associated with breastfed at discharge indicates worse socioeconomic status (e.g., medicaid, WIC food, education level less than high school, etc.). Thus, we hypothesize that women with higher income, better working opportunity and less healthy lifestyles (e.g., smoking) are less likely to choose breastfeeding.

## 5.2 Regression

As Table 2 shows, all five models have similar performance in terms of training and testing MSE, traing and testing R-squared, AIC and BIC. Ridge regression is considered to be the best model because of the shortest training time. As we mentioned in the Methods section, we used gird search to tune the parameters for each model. The hyper-parameter optimization results are shown in the lower table. We were unable to train the best set of RFR due to the time limitation, thus we reported both hyper-parameters and the results of the sub-optimal models. The reported RFR in the table performs a bit worse than other models in terms of R-squared, AIC and BIC but with significantly large training time.

We used RFR to rank feature importance and presented the top 15 most important features. Both continuous variables here have very skewed distribution, which means there are a few dominant features that are very important to predict the outcomes. For example, gestation period is the most important predictor for both outcomes. Infant with less gestation period tends to have lower Apgar score and lower birth weight. Some features such as C-section as delivery method, use of chorioamnionitis, not reporting marital status are the most important predictors for Apgar score, while features including mother's weight, height, BMI, infant's sex and plurality (whether it's single or twin or triplets) are the most important features in predicting birth weight. Mothers' race as black is among the top 15 features for both outcomes. The R-squared is much larger (> 0.50) in predicting birth weight than Apgar score, which means the features in our data can explain more variance of birth weight than the Apgar score.

| Metrics | Train Time | Train Acc | Test Acc | AUC | Prec. | Rec. | F1 score |
|---|---|---|---|---|---|---|---|
| *Living at the report* | | | | | | | |
| GNB | 146s | 0.94 | 0.94 | 0.92 | 0.52 | 0.90 | 0.52 |
| MNB | 34s | 0.82 | 0.81 | 0.88 | 0.51 | 0.82 | 0.46 |
| **Logit** | **203s** | **0.998** | **0.998** | **0.95** | **0.90** | **0.67** | **0.74** |
| **RF** | **182s** | **0.998** | **0.998** | **0.96** | **0.96** | **0.54** | **0.58** |
| *No Abnormal Conditions* | | | | | | | |
| GNB | 165s | 0.85 | 0.85 | 0.76 | 0.65 | 0.69 | 0.67 |
| MNB | 33s | 0.78 | 0.78 | 0.73 | 0.60 | 0.69 | 0.61 |
| **Logit** | **175s** | **0.91** | **0.91** | **0.82** | **0.83** | **0.65** | **0.70** |
| RF | 1826s | 0.91 | 0.91 | 0.82 | 0.85 | 0.64 | 0.69 |
| *No Congenital Conditions* | | | | | | | |
| GNB | 207s | 0.87 | 0.87 | 0.69 | 0.50 | 0.61 | 0.47 |
| MNB | 55s | 0.98 | 0.98 | 0.71 | 0.51 | 0.58 | 0.52 |
| **Logit** | **299s** | **0.997** | **0.997** | **0.75** | **1.0** | **0.55** | **0.60** |
| RF | 383s | 0.997 | 0.997 | 0.73 | 1.0 | 0.55 | 0.58 |
| *Breastfed at Discharge* | | | | | | | |
| GNB | 105s | 0.78 | 0.78 | 0.68 | 0.61 | 0.60 | 0.60 |
| MNB | 23s | 0.69 | 0.69 | 0.67 | 0.59 | 0.63 | 0.59 |
| **Logit** | **218s** | **0.83** | **0.83** | **0.75** | **0.71** | **0.55** | **0.55** |
| **RF** | **1101s** | **0.83** | **0.83** | **0.75** | **0.73** | **0.53** | **0.51** |

**Evaluation metrics of classification**

| Classification | Living | No_Abnorm | No_Congen | BreastFed |
|---|---|---|---|---|
| RFC: max_depth | 10 | 20 | 30 | 30 |
| RFC: min_sample_split | 0.001 | 0.003 | 0.001 | 0.001 |
| RFC: n_estimator | 20 | 100 | 10 | 100 |

**Best set of hyperparameters from grid search**

| | Variable | Coef | P>\|Z\| | 95% CI |
|---|---|---|---|---|
| | *Living at report (N = 2,700,819; Df = 221)* | | | |
| | **weight_gain** | 0.011 | 0.000 | [0.009, 0.013] |
| | **# of prev. cesarean** | -0.148 | 0.004 | [-0.248,-0.048] |
| | **prepragnancy diabetes** | -0.529 | 0.000 | [-0.735, -0.322] |
| | **WIC food** | 0.334 | 0.000 | [0.263, 0.406] |
| | **previsit** | 0.0528 | 0.000 | [0.041, 0.065] |
| | **previsit_sq** | -0.001 | 0.000 | [-0.001, -0.001] |
| | **sex_male** | -0.119 | 0.000 | [-0.178, -0.060] |
| | | | | |
| | *No abnormal conditions (N = 2,702,897; Df = 221)* | | | |
| | **inhospital** | 0.307 | 0.000 | [0.248, 0.367] |
| | **race_AIAN** | 0.107 | 0.003 | [0.036, 0.179] |
| | **race_asian** | 0.094 | 0.001 | [0.039, 0.149] |
| | **hisp_white** | 0.081 | 0.001 | [0.035, 0.127] |
| | **nonhisp_black** | -0.045 | 0.002 | [-0.074, -0.016] |
| | **hisp_black** | -0.165 | 0.000 | [-0.238, -0.093] |
| | **m_height** | 0.0084 | 0.000 | [0.007, 0.010] |
| | **prepregnancy diabetes** | -0.779 | 0.000 | [-0.815, -0.743] |
| | **syphillis infec.** | -0.434 | 0.000 | [-0.553, -0.314] |
| | **Hep C infec.** | -0.434 | 0.000 | [-0.510, -0.358] |
| | **precare** | 0.070 | 0.000 | [0.044, 0.095] |
| | **previsit** | 0.042 | 0.000 | [0.040, 0.045] |
| | **previsit_sq** | -0.001 | 0.000 | [-0.001, -0.001] |
| | **sex_male** | -0.155 | 0.000 | [-0.164, -0.145] |

**Selected coefficients with statistical significance from Logit Models**

Table 1: **Evaluation of classification and important predictors.** The upper left table shows the seven evaluation metrics of classification - Training Time, Training Accuracy, Testing Accuracy, AUC-ROC, Macro Average Precision, Recall and F1 scores. The lower left table shows the hyperparameters used in RFC from grid search of tuning. The right table presents selected coefficients with statistical significance from Logit models predicting living at the report and no abnormal conditions.

| Metrics | Train time | Train MSE | Test MSE | Train R2 | Test R2 | AIC | BIC |
|---|---|---|---|---|---|---|---|
| *5-min Apgar Score* | | | | | | | |
| OLS | 86s | 0.54 | 0.53 | 0.20 | 0.20 | -1680535 | -1677295 |
| **Ridge** | **46s** | **0.54** | **0.53** | **0.20** | **0.20** | **-1680535** | **-1677295** |
| Lasso | 743s | 0.54 | 0.53 | 0.20 | 0.20 | -1678444 | -1675203 |
| ENR | 510s | 0.54 | 0.54 | 0.20 | 0.19 | -1678444 | -1675204 |
| RF | 1618s | 0.54 | 0.54 | 0.19 | 0.19 | -1662630 | -1659389 |
| *Birth weight (kg)* | | | | | | | |
| OLS | 81s | 0.16 | 0.16 | 0.55 | 0.55 | -5003100 | -4999859 |
| **Ridge** | **48s** | **0.16** | **0.16** | **0.55** | **0.55** | **-5003100** | **-4999859** |
| Lasso | 352s | 0.16 | 0.16 | 0.55 | 0.55 | -4960358 | -4957117 |
| ENR | 341s | 0.16 | 0.16 | 0.55 | 0.54 | -4960358 | -4957117 |
| RF | 2487s | 0.17 | 0.17 | 0.51 | 0.51 | -4766148 | -4762907 |

**Evaluation metrics of regression**

| Apgar | Ridge | Lasso | ENR | | RFR | *Reported* | *Best_tuned* |
|---|---|---|---|---|---|---|---|
| *alpha* | 1.0 | 0.0001 | 0.001 | | *max_depth* | 10 | 20 |
| *max_iter* | | 500 | 1000 | | *min_samples_split* | 0.01 | 0.001 |
| *l1_ratio* | | | 0.9 | | *n_estimators* | 20 | 200 |

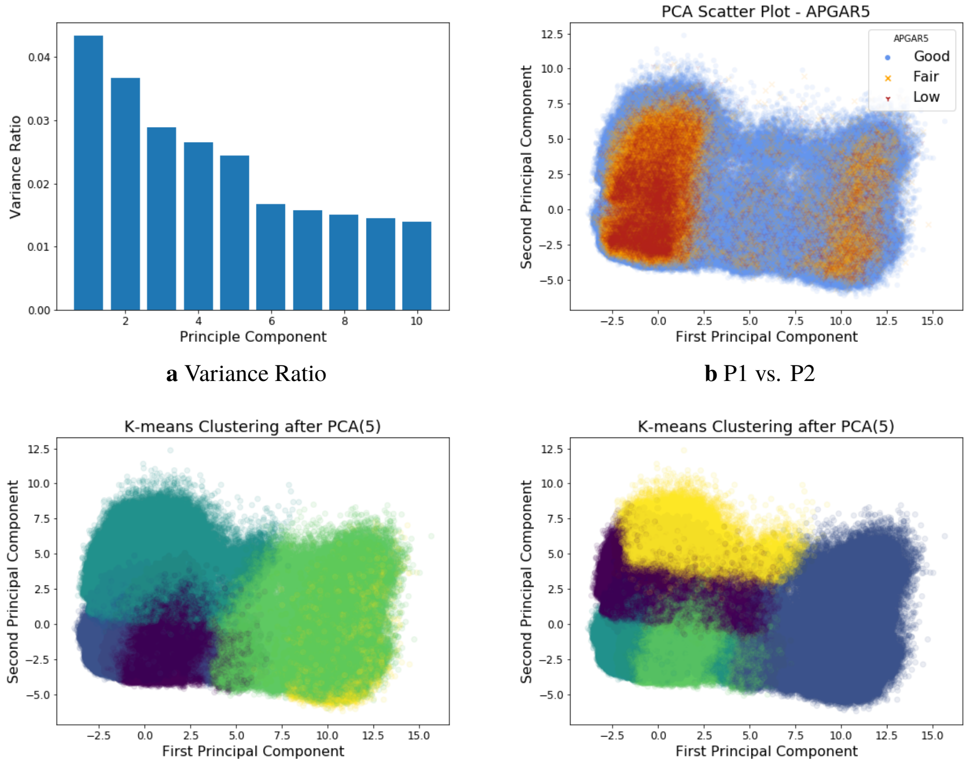
| Weight | Ridge | Lasso | ENR | | RFR | *Reported* | *Best_tuned* |
|---|---|---|---|---|---|---|---|
| *alpha* | 1.0 | 0.001 | 0.001 | | *max_depth* | 10 | 30 |
| *max_iter* | | 500 | 1000 | | *min_samples_split* | 0.01 | 0.001 |
| *l1_ratio* | | | 0.9 | | *n_estimators* | 50 | 200 |

**Best sets of hyperparameters from the grid search**

Table 2: **Evaluation of regression and relevant classification of binary outcome.** The upper table presents seven evaluation metrics of five regression models on two continuous outcomes - 5-min Apgar score and birth weight(kg). The lower table shows the hyperparameters tunes from grid search.

### 5.3 Unsupervised Learning

We tried PCA with different number of components, and decided to set the number of components to ten after which the variance ratio starts to decrease significantly. From the 3a, we can see that the variance ratio is still significant for the tenth component. The results of PCA with ten components are shown in Figure 4. The reconstruction error of PCA is 0.76 in the training set, and 0.77 in the test set. Figure 3b plots the PC1 and PC2 with the three-category 5-minute Apgar label. The plot shows that infants with fair or low Apgar score heavily concentrate in the lower PC1 range and have an other cluster in the higher end, which indicates that the PC1 captures some significant nonlinear risk factors of low Apgar.



**a** Variance Ratio

**b** P1 vs. P2

**c** K-means Clustering after PCA

**d** PCA with K-means labels

Figure 3: **Results of PCA, K-means and K-means Clustering after PCA**

Figure 3c and 3d are scatter plots of the PC1 and PC2, while 3c shows the K-means clustering group membership based on the reduced principal components, whereas 3d is labeled by the K-means clustering group based on the full set of features respectively. Comparing these two figures, we can see that K-means clustering after PCA partially recovers the baseline K-means clustering. That again indicates the good performance of PCA.

We conducted factor analysis with ten components and made similar plots (See Appendix). The results of FA is less interpretable. We compare the log likelihood of PCA and FA with the same number of components. The average log likelihood of PCA is -0.0010 and -0.00024 in the training and test set respectively, whereas the number of FA are 0.00004 and 0.00010.

It is a common practice to build prediction model based on the reduced components. However, in out the data the number of observations is far greater than the number of features, therefore prediction using the reduced components is neither necessary nor meaningful.

## 6 Discussion

We used four classifiers to predict four binary outcomes and used five regression models to predict two continuous outcomes of infant health. Among all the most models, logistic regression classifier and ridge regression are the best models for classification and regression tasks respectively. The accuracy of the classification tasks is consistently over 90% with very large precision, as many of our outcomes are quite rare among the population. Gestation period seem to be the most important features across all the models, which might be an outcome itself. In addition, apart from some important maternal diseases (diabetes, syphilis, Hep C, etc.) we have discovered that in predicting infant health, being black is consistently associated with worse infant health in our models. Although the coefficient is really small compared to the coefficients of the diseases, the statistical significance may indicate that racism still plays a role in persistent socioeconomic gaps, prenatal care, and newborn's health. WIC food program seems to be helpful for low-income women and infants in improving their health. More policy interventions may help close the racial gap.

We think that using the best-tuned hyper-parameters of the random forest after more careful and thorough tuning could achieve the same or probably even better performance than other models. However, given the huge computational costs of a large data and the characteristics of our data ($n \gg p$), simpler models such as the Logistic regression and the Ridge regression can achieve decent performance with significantly less training time. From a lot of failures (dead kernels and endless waiting), we learned how important computational costs are for selecting models to train a big data like this.

To extend our project, we first want to conduct similar analysis on data from previous years to see whether the patterns we found are consistent. We also plan to extend our sample to data from previous years (from 2005s) and put together the annual results to form a time series data (for each day or week). Based on this, we could apply the models for time series analysis and investigate the trends and changes.

# References

[1] Douglas Almond and Lena Edlund. Trivers–willard at birth and one year: evidence from us natality data 1983–2001. *Proceedings of the Royal Society B: Biological Sciences*, 274(1624):2491–2496, 2007.

[2] Hsiang-Yang Chen, Chao-Hua Chuang, Yao-Jung Yang, and Tung-Pi Wu. Exploring the risk factors of preterm birth using data mining. *Expert systems with applications*, 38(5):5384–5387, 2011.

[3] Anthony Costello, Manandhar Dharma, et al. *Improving newborn infant health in developing countries.* Imperial College Press, 2000.

[4] Janet Currie, Michael Greenstone, and Katherine Meckel. Hydraulic fracturing and infant health: New evidence from pennsylvania. *Science advances*, 3(12):e1603021, 2017.

[5] Janet Currie and Hannes Schwandt. The 9/11 dust cloud and pregnancy outcomes: A reconsideration. *Journal of Human Resources*, 51(4):805–831, 2016.

[6] National Center for Health Statistics. *Natality 2017. Public use file. Hyattsville, Maryland: National Center for Health Statistics. Annual internet product.* Available at: `http://www.cdc.gov/nchs/data_access/VitalStatsOnline.htm`, 2018.

[7] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

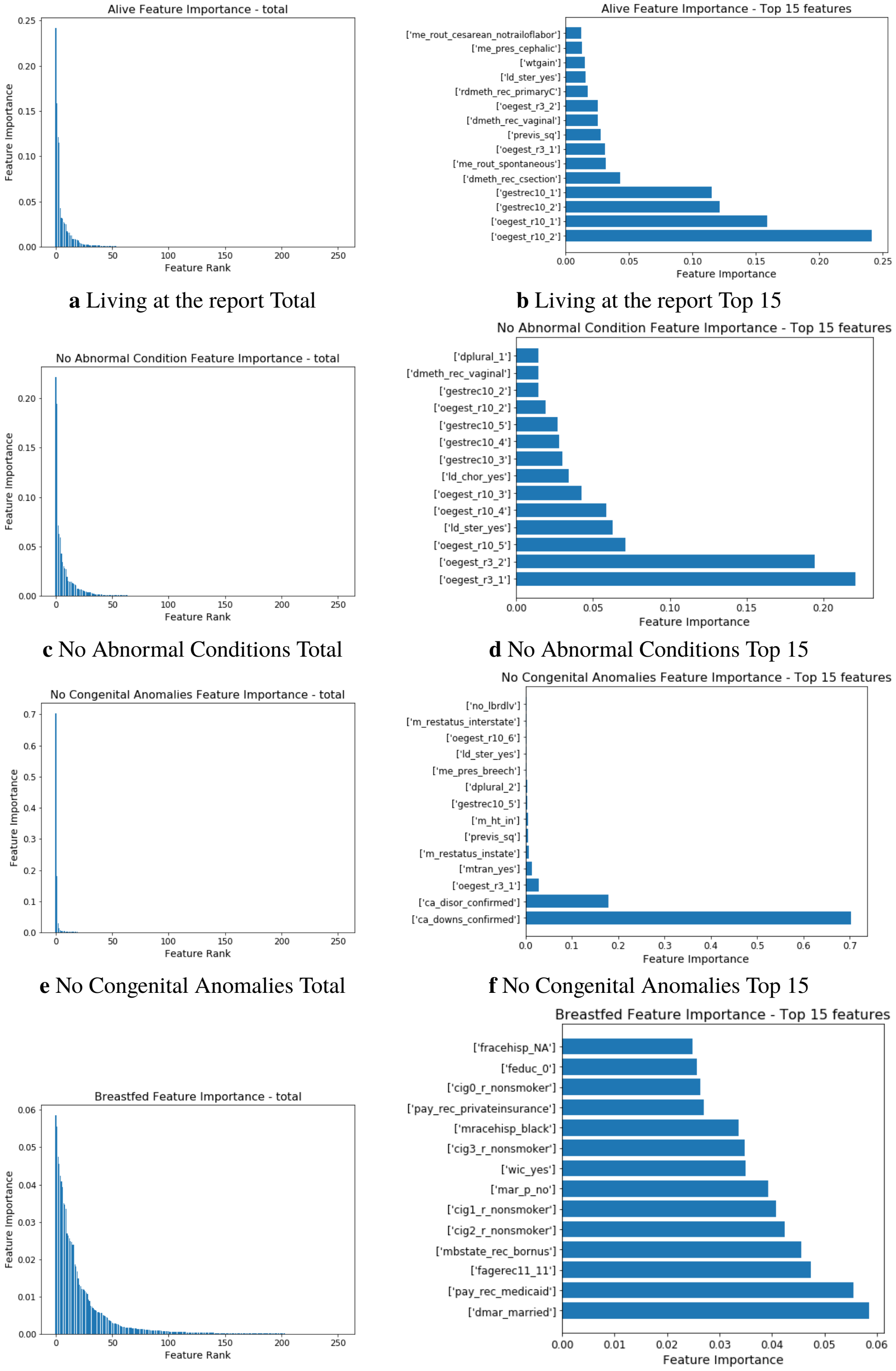
[8] American Academy of Pediatrics et al. The apgar score. *Advances in neonatal care: official journal of the National Association of Neonatal Nurses*, 6(4):220, 2006.

[9] American Academy of Pediatrics et al. Breastfeeding and maternal and infant health outcomes in developed countries. *AAP Grand Rounds*, 18(2):15–16, 2007.

[10] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[11] Corinne A Riddell, Sam Harper, and Jay S Kaufman. Trends in differences in us mortality rates between black and white infants. *JAMA pediatrics*, 171(9):911–913, 2017.

# 7 Appendix



**a** Living at the report Total

**b** Living at the report Top 15

**c** No Abnormal Conditions Total

**d** No Abnormal Conditions Top 15

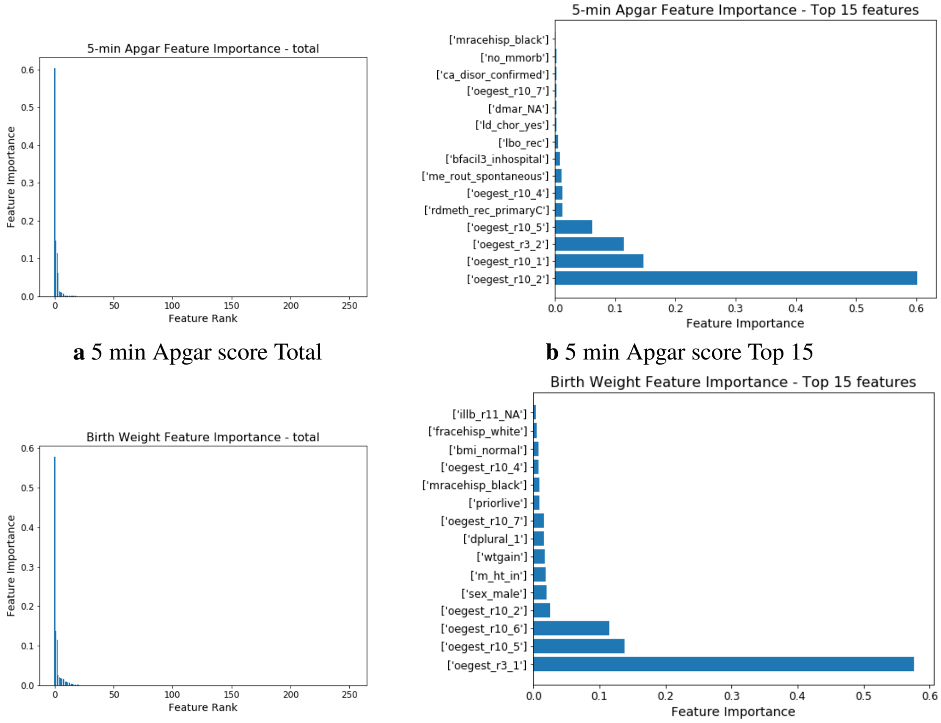
**e** No Congenital Anomalies Total

**f** No Congenital Anomalies Top 15

**g** Breastfed at Discharge Total

**h** Breastfed at Discharge Top 15

Figure 4: **Feature Importance for four binary outcomes**



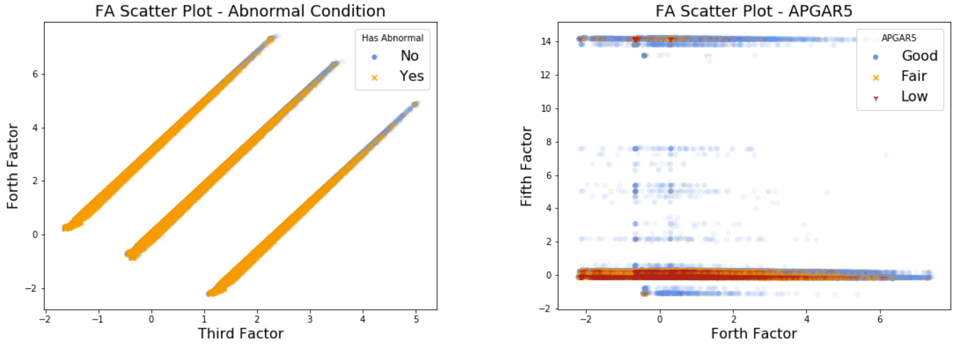
**a** 5 min Apgar score Total

**b** 5 min Apgar score Top 15

**c** Birth Weight Total

**d** Birth Weight Top 15

Figure 5: **Feature Importance for two continuous outcomes**



**a** Factor 3 and 4 Labeled by Abnormal Condition

**b** Factor 4 and 5 Labeled by Apgar Score

Figure 6: **Results from Factor Anlaysis**