Predictive Modeling of the Natality Birth Data

Gayoung Kim, Zhuo Lei, Kagen Quiballo

Introduction

The objective of this project is to find the best model for predicting an infant's Apgar score and interpret each predictor's significance.

The data comes from the 2017 Natality Public Use File from the Centers for Diseases Control and Prevention (CDC). There are over 3 million observations and 281 variables, but we will only analyze the variables that may be significant in predicting an infant's Apgar score.

Methods & Techniques

1. Partitioned data into testing (30%) and training (70%) datasets.
2. Recoded the response variable APGAR5 into a binary variable where score 0-6 is "0" (bad) and 7-10 is "1" (good).
3. Analyzed three different predictive modeling strategies: logistic regression, discriminant analysis, and classification trees.
4. Determined misclassification rates of each confusion matrix.
5. Analyzed variables of importance in context.

Limitations

- Large datasets require more processing power and memory.
- The dataset had 99.7% of high Apgar scores.
  - The proportion of low Apgar scores was proportionally unrepresented and may affect the predictive models.
- The dataset was not cleaned for analysis.
  - Non-categorical data contained nonstandard values as indicators for missing values.
  - Recoded variables were removed to avoid overfitting.

Variables

Response:

Apgar stands for "*Appearance, Pulse, Grimace, Activity, and Respiration*" and is measured once 5 minutes after birth to predict an infant's chances of surviving for the first year of life.

	0 Points	1 Point	2 Points	Points totaled
Activity (muscle tone)	Absent	Arms and legs flexed	Active movement	
Pulse	Absent	Below 100 bpm	Over 100 bpm	
Grimace (reflex irritability)	Flaccid	Some flexion of Extremities	Active motion (sneeze, cough, pull away)	
Appearance (skin color)	Blue, pale	Body pink, Extremities blue	Completely pink	
Respiration	Absent	Slow, irregular	Vigorous cry	

Severely depressed	0-3
Moderately depressed	4-6
Excellent condition	7-10

Pre

Predictive variables included general information about the situation, mother, father, and baby. We chose to focus on variables pertaining to the **mother** and **infant**.

- **General**: *Birthplace, Sex, Labor and delivery*
- **Mother's**: *Age, Race, Marital status, Education, BMI, Weight, Smoking habits, Risk factors, Infections, Interval since last pregnancy, Prenatal care visits, Maternal morbidity, Prior births*
- **Father's**: *Age, Race, Education*
- **Infant's**: *Abnormal conditions, Congenital abnormalities*

Logistic Regression

Analysis of Effects:
32 variables: 6 removed, 26 kept

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
BFACIL	7	25.3995	0.0006
MAGER9	8	16.8848	0.0313
MBRACE	3	52.8741	<.0001
DMAR	1	12.9328	0.0003
MEDUC	8	68.2348	<.0001
FAGEREC11	10	20.6470	0.0237
FRACE6	6	54.1196	<.0001
FEDUC	8	9.8819	0.2734
PRIORLIVE	20	101.1312	<.0001
PRIORDEAD	13	166.7068	<.0001
PRIORTERM	24	62.9570	<.0001
ILLB_R11	10	45.4681	<.0001
ILOP_R11	9	42.0776	<.0001
ILP_R11	10	18.1312	0.0528
PRECARE5	4	87.3952	<.0001
PREVIS_REC	10	517.2931	<.0001
CIG0_R	6	6.6255	0.3569
CIG1_R	6	10.7103	0.0978
CIG2_R	6	11.4671	0.0750
CIG3_R	6	16.8895	0.0097
BMI_R	6	213.4851	<.0001
WTGAIN_REC	5	124.4659	<.0001
NO_RISKS	2	37.9450	<.0001
NO_INFEC	2	12.6985	0.0017
NO_LBRDLV	2	288.3029	<.0001
DMETH_REC	2	120.0321	<.0001
NO_MMORB	2	297.2386	<.0001
GESTREC3	2	7.5567	0.0229
OEGEST_R3	1	6.5786	0.0103
BWTR4	3	6029.9020	<.0001
NO_ABNORM	2	19273.4482	<.0001
NO_CONGEN	1	0.5079	0.4760

c-value = 0.853 (close to 1)

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	85.3	Somers' D	0.705
Percent Discordant	14.7	Gamma	0.705
Percent Tied	0.0	Tau-a	0.029
Pairs	29161219318	c	0.853

Goodness of Fit Test p-value < 0.05
the model is a good fit.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
464.4375	8	<.0001

Our model correctly predicted ~**98%** of bad-good Apgar scores.

Table of F_APGAR_Y by I_APGAR_Y			
F_APGAR_Y(From: APGAR_Y)	I_APGAR_Y(Into: APGAR_Y)		
	0	1	Total
0	480 0.09 4.55 40.85	10064 1.98 95.45 1.98	10544 2.07
1	695 0.14 0.14 59.15	497731 97.79 99.86 98.02	498426 97.93
Total	1175 0.23	507795 99.77	508970 100.00

Discriminant Analysis

Summary of Stepwise Selection:
10 num. variables: 2 removed, 8 kept

Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F
1	1	PREVIS		0.0024	6114.25	<.0001
2	2	BMI		0.0010	2491.98	<.0001
3	3	PRIORLIVE		0.0004	1089.11	<.0001
4	4	PRIORDEAD		0.0002	550.83	<.0001
5	5	WTGAIN		0.0002	491.03	<.0001
6	6	PRIORTERM		0.0002	456.15	<.0001
7	7	CIG_0		0.0001	300.22	<.0001
8	8	M_HT_IN		0.0000	12.63	0.0004
9	9	PWGT_R		0.0000	9.14	0.0025
10	8		BMI	0.0000	1.22	0.2686
11	9	DWGT_R		0.0000	7.84	0.0051
12	8		WTGAIN	0.0000	0.60	0.4398
13	9	CIG_2		0.0000	7.39	0.0066
14	10	CIG_3		0.0000	25.99	<.0001

0.05 ☐ use *quadratic discriminant function*.

Manova Test p-values < 0.05
☐ mean for different classes across different predictors are *significantly different*.

Multivariate Statistics and Exact F Statistics S=1 M=4.5 N=1273996.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.99541644	1066.61	11	2.55E6	<.0001
Pillai's Trace	0.00458356	1066.61	11	2.55E6	<.0001
Hotelling-Lawley Trace	0.00460467	1066.61	11	2.55E6	<.0001
Roy's Greatest Root	0.00460467	1066.61	11	2.55E6	<.0001

Use priors **proportional** because the proportions for bad Apgar_Y to good Apgar_Y are not equal.

Our model correctly predicted ~**94%** of bad-good Apgar scores.

Number of Observations and Percent Classified into APGAR_Y			
From APGAR_Y	0	1	Total
0	1631 7.81	19244 92.19	20875 100.00
1	47022 4.39	1024235 95.61	1071257 100.00
Total	48653 4.45	1043479 95.55	1092132 100.00
Priors	0.01915	0.98085	

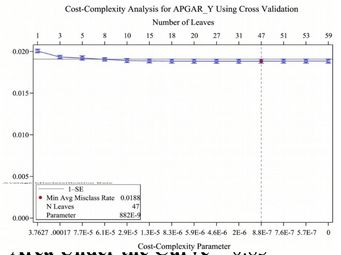
Error Count Estimates for APGAR_Y			
	0	1	Total
Rate	0.9219	0.0439	0.0607
Priors	0.0192	0.9808	

Classification Trees

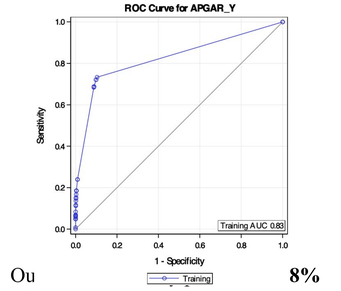
Variables Importance:
32 variables: 21 removed, 11 kept

Variable Importance			
	Training		
Variable	Relative	Importance	Count
DBWT	1.0000	87.3799	27
NO_ABNORM	0.8396	73.3644	1
GESTREC3	0.1847	16.1414	3
NO_CONGEN	0.1425	12.4560	2
DPLURAL	0.0915	7.9925	1
WTGAIN	0.0763	6.6638	1
PREVIS	0.0657	5.7430	2
ATTEND	0.0490	4.2775	1
BMI	0.0273	2.3860	1
ILLB_R11	0.0261	2.2815	1
FAGEREC11	0.0259	2.2659	1

Cross-Complexity Analysis:
47 leaves has the minimum average misclassification rate of 0.0188.
We will use this selected tree.



Area Under the Curve 0.83



Ou **8%** of bad-good Apgar scores.

Confusion Matrices				
		Predicted		
	Actual	0	1	Error Rate
Cross Validation	0	4520	40087	0.8987
	1	1785	2178082	0.0008

Summary of Best Model

Correct Classification Rate	
Logistic Regression	97.9%
Discriminant Analysis	93.9%
Classification Tree	98.1%

Area Under the Curve	
Logistic Regression	0.853 (c-value)
Classification Tree	0.83 (AUC)

Type I Error Rate (False-Positive)	
Logistic Regression	59.1%
Discriminant Analysis	96.6%
Classification Tree	28.3%

Conclusion: Both LR and CT have a high classification rate. Although LR has a slightly higher AUC, it also has a significantly greater type I error rate. Hence, **classification tree** model is the best predictive model for Apgar scores.

Variables	Healthy	Unhealthy
1. Birthweight	5 lbs 8oz – 8 lbs 13 oz	< 5 lbs 8oz or > 8 lbs 13 oz
2. No Abnormal Conditions	No seizures, antibiotics required, ventilation required, etc.	Has abnormal condition or requires treatment
3. Gestational Age	38-42 weeks	< 38 weeks or > 42 weeks

Future Directions

- Dataset is very likely to contain variables that are *correlated* with each other. A **principal component analysis** will allow us to detect interesting features and underlying patterns such as grouping behaviors among variables.
- For infants with bad Apgar5 scores, we could analyze their **Apgar 10 scores** to see *how they changed*.
- For variables that only test for the **general presence** of an abnormality or condition, we could run more specific models on *which specific abnormalities and conditions* are significant predictors of Apgar score.