Predictive Modeling of the Natality Birth Data

Gayoung Kim, Zhuo Lei, Kagen Quiballo

Introduction & Methods

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 underrepresented 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

Sippel, Robert. "Apgar Scoring System." *EMS World.* 18 Jan 2012. https://www.emsworld.com/article/10615556/ems-recap-apgar-scoring

Predictors

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

Predictive Models

Logistic Regression

procedure: PROC LOGISTIC

ANALYSIS OF EFFECTS:

32 variables: 6 removed, 26 kept

C-VALUE:

$= 0.853$ (close to 1)

GOODNESS OF FIT TEST:

p-value $= <$ 0.05 *the model is NOT a good fit.*

CONFUSION MATRIX:

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

Discriminant Analysis

Procedure: PROC DISCRIM

STEPWISE SELECTION (proc stepdisc):
10 numeric variables: 2 removed, 8 kept

TEST OF HOMOGENEITY:
p-value < 0.05 *use quadratic discriminant function*

MANOVA TEST:
p-values < 0.05 *mean for different classes across different predictors are significantly different*

PRIORS PROPORTIONAL:
the proportions for bad Apgar_Y to good Apgar_Y are not equal

CONFUSION MATRIX:
Our model correctly predicted ~94% of bad-good Apgar scores

Classification Trees

Procedure: PROC HPSPLIT

VARIABLE IMPORTANCE:

32 variables: 21 removed, 11 kept

CROSS -COMPLEXITY ANALYSIS:

47 leaves has the minimum average misclassification rate of 0.0188

AREA UNDER THE CURVE:

0.83 (close to 1)

CONFUSION MATRIX:

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

Conclusion

Summary of Best Model

Predictive Model	Correct Classification Rate	Area Under the Curve	Type I Error Rate (False-Positive)
Logistic Regression	97.9%	0.853 (c-value)	59.1% (G.O.F.)
Discriminant Analysis	93.9%	-------	96.6%
Classification Tree	98.1%	0.83 (AUC)	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.

Analysis of Variables

Most Significant Variables based on Classification Tree Model

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

Conclusion: It is highly important for an infant to maintain a healthy weight to predict high likelihood to survive in the first **year of the life.**

Future Directions

Future Directions

1. 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.

2. For infants with bad Apgar5 scores, we could analyze their **Apgar 10 scores** to see *how they changed*.

3. 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.

THANK YOU

QUESTIONS