**Predicting Infant Mortality Using Logistic Regression**

**Capstone Project 1 (Milestone Report)**

**Springboard Data Science Career Track**

**By**

**Shishay Bisrat**

**September 2017**

**Predicting Infant Mortality Using Logistic Regression**

1. **Define the problem**

One of the most heart-breaking experiences in life is losing an infant to untimely death. The National Center for Health Statistics (NCHS) collects and performs statistical analysis of birth and infant death in the US and its territories on a yearly basis. Most of the infant deaths are linked to the birth files providing a publicly available good data set for anyone interested in detailed investigation of infant death in the country. In this project Logistic Regression is used to build a model that predicts infant mortality using routinely collected infant health and birth data.

1. **General information and potential client**

Infant mortality is defined as death of children who have lived less than one year. Studies have shown that lack of proper medical care during pregnancy, delivery, and immediately after delivery, strongly increase the mortality risk of an infant. The NCHS routinely compiles data it collects at each stage of the life of an infant, including: maternal risk factors, pregnancy-related health data, delivery procedures and complications, and anomalies at birth.

The purpose of this project is to build a model trained on such data to predict infant mortality, evaluate feature importance, and make recommendations based on the project’s results. The results of this project would help different organizations and personnel including health care professionals, researchers, insurance companies, and parents to make informed decisions.

**3. Dataset and Data Wrangling**

The data used in this project (a.k.a. birth cohort data) includes natality (birth data) and fatality (death data) of infants born in the year 2008, along with a document that describes the data content and type. The birth cohort data for 2008 consists of infant deaths that occurred in 2008 or 2009 linked to births in 2008. The data also includes a separate file that includes infant deaths, and unlinked file, which consists of infant deaths that had not been linked to a corresponding record in the natality file.

The document that describes the data, originally in PDF, was converted into usable format in two steps. First, Tabula (Aristaran and Tigas, 2013), a software service, was used to convert the guide document from pdf to TSV (tab separated values file) format. The TSV file was then reformatted using a Python code. The original data set did not have column names and it required a second step of writing Python code to extract the field names from the guide document.

Some columns did not have any values and were removed at the beginning of data wrangling. These include features such as county and state of residence of infant’s mother. The guide document lists the valid values for each feature. I wrote a Python code that builds a dictionary of the valid values for each column. Any value outside the list, invalid value, was converted to NaN (Python’s denotation of the concept “Not A Number”). After removing all rows that consisted of NaN values the dataset consisted of in 1,569,762 rows (records) and 102 columns (features).

**4. Other potential datasets**

The data wrangling process produced a clean data set with 1,569,762 rows (records) and 102 columns (features). This is only from the year 2008. I can use more data that spans from 1994 to 2010. This is huge dataset, each year containing 3-4 million records.

**5. Initial findings**

The data attributes included and used in this project are:

**Infant Information:** Sex (SEX), Gestation (COMBGEST), Clinical Gestation Estimate (ESTGEST), Birth Weight (BRTHWGT), Plurality (DPLURAL), Apgar Score (APGAR5), Birth Place Revised (BFACIL), Birth Place (UBFACIL).

**Mother Information:** Mother’s Age (MAGER), Mother’s Bridged Race (MBRACE), Mother’s Race Recode (MRACEREC), Mother’s Hispanic Origin (MRACEHISP, UMHISP), Mother’s Education (MEDUC), Marital Status (MAR), Residence Status (RESTATUS).

**Father Information:** Father’s age (FAGECOMB), Father’s Bridged Race (FBRACE), Father’s Hispanic Origin (FRACEHISP, UFHISP).

**Pregnancy Information**: Total Birth Order (TBO), Live Birth Order (LBO), Number of Prenatal Visits (UPREVIS), Month Prenatal Care Began (PRECARE), Weight Gain (WTGAIN).

**Delivery Methods**: Attempted Forceps (ME\_ATTF), Attempted Vacuum (ME\_ATTV), Fetal Presentation (ME\_PRES), Route and Method of Delivery (ME\_ROUT), Trial of Labor Attempted (ME\_TRIAL), Forceps (UME\_FORCP), Vacuum (UME\_VAC), Delivery Method Recode (RDMETH\_REC), Delivery Method Recode (DMETH\_REC), Attendant (ATTEND).

**Risk Factors**: Pre-pregnancy Diabetes (RF\_DIAB), Gestational Diabetes (RF\_GEST), Pre-pregnancy Hypertension (RF\_PHYP), Gestational Hypertension (RF\_GHYP), Eclampsia (RF\_ECLAM), Previous Preterm Birth (RF\_PPTERM), Poor Pregnancy Outcome (RF\_PPOUTC), Previous Cesarean Deliveries (RF\_CESAR), Number of Previous Cesarean Deliveries (RF\_CESARN).

**Risk Factors in this Pregnancy**: Diabetes (URF\_DIAB), Chronic Hypertension (URF\_CHYPER), Pregnancy-Associated Hypertension (URF\_PHYPER), Eclampsia (URF\_ECLAM).

**Obstetric Procedures**: Cervical Cerclage (OP\_CERV), Tocolysis (OP\_TOCOL), Successful External Cephalic Version (OP\_ECVS), Failed External Cephalic Version (OP\_ECVF), Induction of Labor (UOP\_INDUC), Tocolysis (UOP\_TOCOL).

**Onset of Labor**: Premature Rupture of Membrane (ON\_RUPTR), Precipitous Labor (ON\_PRECIP), Prolonged Labor (ON\_PROL).

**Characteristics of Labor and Delivery**: Induction of Labor (LD\_INDL), Augmentation of Labor (LD\_AUGM), Non-Vertex Presentation (LD\_NVPR), Steroids (LD\_STER), Antibiotics (LD\_ANTI), Chorioamnionitis (LD\_CHOR), Meconium Staining (LD\_MECS), Fetal Intolerance (LD\_FINT), LD\_ANES (Anesthesia).

**Complications of Labor and Delivery**: Meconium (ULD\_MECO), Precipitous Labor (ULD\_PRECIP), Breech (ULD\_BREECH).

**Tobacco Use**: Cigarette Recode (CIG\_REC), Cigarettes First Trimester (CIG\_1), Cigarettes Second Trimester (CIG\_2), Cigarettes Third Trimester (CIG\_3).

**Abnormal Conditions of Newborn**: Assisted Ventilation (AB\_AVEN1), Assisted Ventilation > 6hrs (AB\_AVEN6), Admission to NICU (AB\_NICU), Surfactant (AB\_SURF), Antibiotics (AB\_ANTI), Seizures (AB\_SEIZ), Birth Injury (AB\_BINJ).

**Congenital Anomalies of the Newborn**: Anencephaly (CA\_ANEN), Meningomyelocele/Spina Bifida (CA\_MNSB), Cyanotic Congenital Heart Disease (CA\_CCHD), Congenital Diaphragmatic Hernia (CA\_CDH), Omphlocele (CA\_OMPH), Gastroschisis (CA\_GAST), Limb Reduction Defect (CA\_LIMB), Cleft Lip w/ or w/o Cleft Palate (CA\_CLEFT), Cleft Palate Alone (CA\_CLPAL), Downs Syndrome (CA\_DOWN), Suspected Chromosomal Disorder (CA\_DISOR), Hypospadias (CA\_HYPO), Anencephalus (UCA\_ANEN), Spina Bifida/ Meningocele (UCA\_SPINA), Omphalocele/ Gastoschisis (UCA\_OMPHA), Cleft Lip/ Palate (UCA\_CELFTLP), UCA\_DOWNS (Downs Syndrome).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Attribute | Mean | St. Dev. | Min | 25% | 50% | 75% | Max |
| APGAR5 | 8.8 | 7.6 | 0.00 | 9.0 | 9.0 | 9.0 | 10.0 |
| BRTHWGT | 3,300.7 | 568.3 | 229 | 3,005 | 3,331 | 3,657 | 8,136 |
| CIG\_1 | 0.9 | 3.9 | 0 | 0 | 0 | 0 | 98 |
| CIG\_2 | 0.7 | 3.2 | 0 | 0 | 0 | 0 | 98 |
| CIG\_3 | 0.7 | 3.1 | 0 | 0 | 0 | 0 | 98 |
| COMBGEST | 38.7 | 2.3 | 17 | 38 | 39 | 40 | 47 |
| DPLURAL | 1.0 | 0.2 | 1 | 1 | 1 | 1 | 5 |
| ESTGEST | 38.5 | 2.0 | 3 | 38 | 39 | 40 | 50 |
| FAGECOMB | 30.5 | 7.0 | 10 | 25 | 30 | 35 | 86 |
| LBO | 2.1 | 1.2 | 1.0 | 1.0 | 2.0 | 3.0 | 8.0 |
| MAGER | 28.0 | 6.0 | 12.0 | 23.0 | 28.0 | 32.0 | 50.0 |
| MEDUC | 4.1 | 1.8 | 1.0 | 3.0 | 4.0 | 6.0 | 8.0 |
| PRECARE | 3.0 | 1.5 | 0.0 | 2.0 | 3.0 | 3.0 | 10.0 |
| RF\_CESARN | 0.2 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 10.0 |
| TBO | 2.4 | 1.5 | 1.0 | 1.0 | 2.0 | 3.0 | 8.0 |
| UPREVIS | 11.3 | 3.8 | 0.0 | 9.0 | 11.0 | 13.0 | 49.0 |
| WTGAIN | 33.7 | 19.4 | 0.0 | 22.0 | 31.0 | 41.0 | 99.0 |

**Table 1**: Descriptive statistics of the numerical attributes

Direct inspection of Table 1 reveals that there does not appear to be any anomaly from the descriptive statistics of the numerical attributes. It is better to look at the distribution of individual attributes to see “anything interesting”. Notice, however, that the mean birth weight and estimated gestation period are 3,300.7 grams and 38.5 weeks, respectively.

**4.1 Investigating Correlations Between APGAR Score and Infant Mortality**

APGAR score is a measure of the physical condition of a newborn infant. The APGAR score from the mortality data shows a bimodal distribution with maximum counts at APGAR score of 01 and 09 (Figure 1). Typically, it is not expected for an infant of high APGAR score to die.

**Figure 1**: Apgar score has a bimodal distribution in the infant mortality dataset. The causes of death for infants with high APGAR score are accidents or medical anomalies.

To understand the anomaly, I investigated the causes of death of infants with high APGAR score. The three major causes of death for infants with high APGAR score are sudden infant death syndrome, R99 (symptoms, signs and abnormal clinical and laboratory findings not classified anywhere), and accidental suffocation or strangulation in bed.

**4.2 Investigating relationship between gestation period and infant mortality**

The gestation period, the length of pregnancy, is critical to the development of an embryo/fetus. In humans, the gestational age is about 40 weeks although it is common to see births at gestational age of 37 to 42 weeks.



**Figure 2**: Gestation period has a bimodal distribution in the infant mortality dataset. There is a minor peak at 23 weeks in the mortality dataset although the average gestational age for humans is 40 weeks.

The bar chart for the gestation period (Figure 2) shows the infant mortality has a bimodal distribution with two peaks at 23 and 39 weeks. For the infant non-mortality dataset, the distribution of gestation is period is normal with one strong peak at 39 weeks. This suggests the importance of gestation period in the mortality rate of an infant.

**4.3 Summary of EDA**

The descriptive statistics of numerical attributes shows values that would be categorized as normal for each attribute. A deeper investigation with the help of visualization for APGAR score reveals some interesting insights. The APGAR score histogram chart shows a minor peak at high APGAR score values for infant death set. Normally, high APGAR score is expected to correlate with low infant death rate. However, sudden infant death syndrome, suffocation, and unclassified infant death anomalies are the main causes of death for infants with high APGAR score. The bar charts for anomalies show that admission to ICU has a significant weight in predicting the likelihood of infant death.

The bar chart for gestation period show a single peaked normal distribution for the infant non-mortality dataset while exhibiting bimodal distribution for infant mortality dataset.

**5. Classification Using Logistic Regression**

**5.1 Building a Base Model**

I used logistic regression to build a base model and the general performance of the classifier was evaluated using known performance metrics. The objective of the classifier is to predict the risk (as a probability) of mortality of an infant using information collected about the mother, father, and child during the mother’s pregnancy and child delivery. The procedures below were followed to build the model.

1. Separate the attributes into categorical (nominal), numerical (noncategorical), and binary. The dataset consists of 25 categorical, 17 numerical and 60 binary features.
2. Encode the categorical attributes, as most machine algorithms require numeric inputs and outputs for efficient implementation.
3. Specify predictor features and target features. The target feature in this project is the mortality or non-mortality of an infant. The mortality and non-mortality of an infant are assigned a positive and a negative label, respectively. The predictor features increased to 917 after encoding the categorical features using dummy variables.
4. Split the dataset into training and test datasets (90% training and 10% test data set)

keeping track of the positive to negative label ratio in both data sets.

1. Use Grid Search and Cross-Validation (using k-fold cross validation with a k= 5) and determine the optimum regularization parameter.
2. Train the best algorithm on the training set and test the algorithm on the test set.
3. Evaluate the classifier using performance metrics. Use classification report (from package imblearn; Lemaire et. al., 2017) to evaluate a classifier trained using the original data.

**5.1.1 Performance Results**

**Table 2**: Performance metrics of the classifier from sklearn’s classification report (using imbalanced data)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Class | Accuracy | Prec. | Recall | Specificity | F1 | Geom. Mean\* | Iba\* | Support |
| Train Set | 0 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1406613 |
| 1 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 6172 |
| Avg\* | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1412785 |
| Test Set | 0 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 156318 |
| 1 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 659 |
| Avg\* | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 156977 |

**\*Avg = Average; Prec. = Precision; Geom. Mean = Geometric Mean; Iba = Index Balanced Accuracy**

The performance of the classifier, using sklearn’s imbalanced classification report to take the effect of data imbalance into consideration, is shown in Table 2. A quick look at the table raises suspicion, as even the best classifiers are not expected to perform perfectly.

The plan is to continue investigating the effect of data imbalance and use techniques (such as resampling) to build a reliable model.