**Predicting Infant Mortality Using Logistic Regression**

**Capstone Project 1**

**Springboard Data Science Career Track**

**By**

**Shishay Bisrat**

**January 2018**

**Predicting Infant Mortality Using Logistic Regression**

1. **Abstract**

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 paper Logistic Regression is used to build a model that predicts infant mortality using routinely collected infant health and birth data. The infant mortality/non-mortality data is inherently imbalanced and require data resampling in a preprocessing stage. Experimenting with different under-sampling and oversampling methods indicated that removing Tomek Links followed by under-sampling gives the best results. Feature importance analysis shows that admission to Intensive Care Unit, Cyanotic Congenital Heart Disease, Spina Bifida/ Meningocele are the three major attributes with high positive coefficients and hence high effect on infant mortality (positive class). Apgar score has a high negative coefficient showing that it has high effect size on infant non-mortality (negative class).

1. **General Information**

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. Exploratory Data Analysis (EDA)**

The data wrangling process produced a clean data set with 1,569,762 rows (records) and 102 columns (features). The attributes of these features and their proportion are shown in Figure 1.

****

**Figure 1**: Most of the data points (~59%) in the dataset are binary, answering yes or no questions, such as the presence of anomaly. The pie chart shows the proportion of binary, categorical, and numeric data in the dataset.

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 2). Typically, it is not expected for an infant of high APGAR score to die.

**Figure 2**: 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 3**: 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 3) 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 Investigating relationship between birth weight and infant mortality**

Most healthy babies born at the normal gestation period (37 to 40 weeks) weigh between 2500 and 4000 grams. In addition to gestation period, weight of a newborn can be affected by size of parents, gender, health of mother, nutrition, etc.

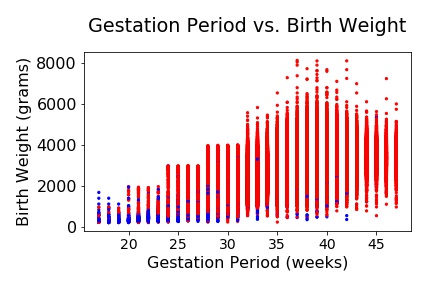


**Figure 4**: The birth weight of newborn babies shows a different distribution for the infant mortality and infant non-mortality dataset. The major peak for the infant mortality data set is 750 grams. The birth weight for the non-mortality dataset has normal distribution with a single major peak at 3000 grams.

The birth weight of the newborn from the infant non-mortality data is normally distributed with mean and median weights of 3,307 grams and 3,335 grams respectively (Figure 4). The birth weight of newborn babies from the infant mortality dataset has two peaks with a major peak at 750 grams. The mean and median weight for the mortality dataset is 1,842 grams and 1,786 grams respectively.

**4.4 Investigating relationship between birth weight and infant mortality**

Generally, the shorter the gestation period the lower the birth weight of a newborn baby is. This is not surprising, as on average, the weight of a fetus progresses from 500 grams (23 weeks), to 1 kg (28 weeks), to 2.1 kg (34 weeks), and 3.5 kg (40 weeks). The bar chart in Figure 5 shows the infant mortality dataset is predominantly consisted of low birth weight and short gestation period.



**Figure 5**: The birth weight is significantly influenced by the gestation period. The shorter the gestation period, the lower the birth weight is. The figure shows infant mortality is generally related to lower birth weight and shorter gestation period (blue circles).

**4.5 Visualizing the proportion of abnormality in infant mortality/natality datasets**

Anomalies or complications during delivery lead to the administration of various medical procedures, such as assisted ventilation, antibiotics, and admission to ICU. Separate plots for infant mortality and infant non-mortality dataset (Figure 6) show the proportion of such medical procedures were higher for the infant mortality dataset.

|  |
| --- |
|  |
|  |
|  |
|  |

**Figure 6**: Some anomalies such as admission to ICU and administering assisted ventilation for more than six hours appear to be highly correlated with infant mortality (codified as “1”).

**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 charts and histograms for gestation period and birth weight show a single peaked normal distribution for the infant non-mortality dataset while exhibiting bimodal or skewed distributions 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. An imbalance in data might explain near perfect scores in accuracy, but not the other metrics such as recall. Rather, perfect scores might indicate the presence of features that act as proxies to the target labels. For example, in the common instance of using height and weight of individuals to classify them to males and females, a third attribute (such as level of testosterone) may serve as proxy to the output labels increasing the overall performance of the classifier but significantly diminishing the weight of the height and weight. In this project, four attributes, ICD code for infant death (UC0D), cause of infant death recode (UCODR130), place of death and manner of death are linked to the positive class (infant death) only. In the dataset, these attributes have a range of values for infant death records but one constant value for surviving infants. Therefore, these four attributes, if they have any value different from the constant for surviving infants, can act as proxies for the positive class. This proposition is supported by their high coefficients observed in feature importance analysis. Confusion matrix, precision-recall curve, and ROC for both training and test sets also show the unsatisfactory results of using the imbalanced data to predict infant mortality (see Figure 7).

Therefore, the four features were removed from further consideration for two main reasons. First, the information about these attributes are collected at death and are not useful for the classifier, which purports to use data from pregnancy and delivery, to predict mortality or non-mortality of an infant. Second, they act as proxy features, practically invalidating the necessity for the rest of the attributes.

|  |  |  |
| --- | --- | --- |
|  | **Train Set** | **Test Set** |
| **Confusion Matrix** |  |  |
| **Precision-Recall Curve** |  |  |
| **ROC** |  |  |

**Figure 7**: Confusion matrix, precision-recall curve, and ROC curve for the training set (left column) and test set (right column) indicate the unsatisfactory results of the classifier trained on imbalanced data.

To isolate the effect of data imbalance, the proxy features were dropped from the predictors, and a new model was built. Removing the proxy features reduced the number of categorical features to 21, while the numerical and binary features remained 17 and 60 respectively. The total number of predictor features is 192 after encoding the categorical features. The performance metrics for both the train and test sets are tabulated below.

**Table 3**: Performance metrics of the logistic regression classifier after removing proxy features.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Class | Accuracy | Prec.\* | Recall | Specificity | F1 | Geom. Mean\* | Iba\* | Support |
| Train Set | 0 | 1.00 | 1.00 | 1.00 | 0.30 | 1.00 | 0.85 | 0.74 | 1406650 |
| 1 | 0.30 | 0.73 | 0.30 | 1.00 | 0.43 | 0.85 | 0.71 | 6135 |
| Avg\* | 1.00 | 1.00 | 1.00 | 0.30 | 1.00 | 0.85 | 0.74 | 1412785 |
| Test Set | 0 | 1.00 | 1.00 | 1.00 | 0.32 | 1.00 | 0.87 | 0.77 | 156281 |
| 1 | 0.32 | 0.76 | 0.32 | 1.00 | 0.45 | 0.87 | 0.74 | 696 |
| Avg\* | 1.00 | 1.00 | 1.00 | 0.33 | 1.00 | 0.87 | 0.77 | 156977 |

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

Results from the above table merit detailed discussion in the presence of an imbalanced data. The salient interpretations from the performance metrics table, confusion matrix, Precision-Recall Curve, and ROC are detailed below.

**Average vs Individual Class Performance:** When working with imbalanced data taking the performance of the classifier as an average of both classes can be misleading. For example, if we look at the training set, the average recall is 1.00 although the recall for the positive class (label 1) is 0.30. Simply looking at the average recall (1.00) of the train set would hide the poor performance with respect to the positive class. Therefore, it is recommended to evaluate the metrics with respect to each class it is designed to evaluate. The recall and specificity measure the performance of the classifier with respect to the positive and negative class respectively.

**Accuracy:** The average accuracy for both the train set and test sets is 0.996. The accuracy for the positive class (equivalent to recall) is only 0.30 in the train set and 0.32 in the test set. Accuracy uses count of errors and a simple threshold of 0.5 in classification decisions making it unreliable when evaluating the performance of a machine-learning algorithm for an imbalanced data. In this project, the ratio of negative to positive labels is 228:1. Therefore, simply classifying a data point to the majority class (naïve-classifier) gives an accuracy of 99.6%. However, simply classifying to the majority class (non-mortality) beats the purpose of the classifier, as we are more interested in the infant mortality. Given a set of data, the classifier would predict infant mortality correctly, only 32% of the time. This is likely below the minimum threshold in most cases—which is typically defined as 50%, or “flip the coin” performance.

**Precision**: The average precision is 1.00 for both train and test sets. The precision for the positive class is 0.73 for the train set and 0.76 for the test set. This normally would indicate that the classifier is correct 76% of the time when it predicts an infant mortality. However, the precision can be high in an imbalanced class where the True Negative Rate (TNR) is very high, which reduces the false positives, which in turn increases the precision.

**Recall (Sensitivity)**: The recall is 0.30 for the train class and 0.32 for the test class. This indicates that there the classifier will correctly predict 32% of the total infant mortality for a given year. This is very low prediction rate suggesting the need to explore more avenues for improving the performance of the classifier.

**Specificity (TNR)**: The specificity is 1.00 due to the high proportion of negative labels in both train and test sets. As discussed previously, the high specificity resulted in the high precision and accuracy.

**F1-score**: F1 score is a measure of accuracy and is calculated by finding the harmonic mean of precision and recall. The F1 score is 0.43 and 0.45 for the train and test set respectively. The low F1 score is mainly due to the low recall (0.30 and 0.32 for the train and test set respectively).

**Geometric mean:** The geometric mean, as its name implies, is the square root of the product of the specificity and recall. Combining the two metrics help account for balancing of the data set. The geometric mean is 0.85 and 0.87 for the train and test sets respectively. The high geometric mean value is due to the high specificity value.

The performance metrics above show that the classifier is not adequate for a reliable prediction. This is not unexpected as the classifier is trained on a severely imbalanced data set. The best way to deal with such data set is use machine learning tools specifically designed for imbalanced data.

**5. 2 Classification of imbalanced data**

The dataset of infant mortality/non-mortality is inherently highly imbalanced with only a fraction of infants of all born babies dying as infants. This leads to what is commonly called a “class imbalance classification problem” where the majority class predominantly influences the classifier. In this section, the class imbalance in the data set, the irrelevance of accuracy as an evaluation yardstick, and tools to overcome class imbalance-related problems are discussed in detail.

**5.2.1 Data Imbalance**

Many real world classification problems deal with classes that have disproportionate number of members, for instance: the percentage of financial fraud, infant fatality, and product defects are all small compared to the overall transactions, natality rate, and product output respectively.

The infant mortality/non-mortality dataset has severe class imbalance as shown in Figure 8. The current project is limited to the year 2008, where the original mortality/non-mortality is 27,895 and 4,255,188 infants respectively. After data wrangling, the number of mortality/non-mortality is reduced to 1,562,931 and 6,831 respectively.



**Figure 8**: Infant mortality/non-mortality in the US for the years 2002-2009 shows the inherent imbalance in the classification problem.

**5.3 Balancing Data Using Resampling Methods**

One way to counteract the issue of data imbalance in classification problems is to balance the data. There are two major methods of balancing data: under-sampling and over-sampling, which are collectively referred to as resampling methods.

Table 4 shows a summary of the results obtained after applying several resampling methods. A discussion of these methods and associated results follow.

**5.3. 1 Under-sampling Methods**

Under-sampling methods employ algorithms that sample subsets of the majority class to balance the data set. I used Random Under-Sampling (RUS), RUS/TomekLinks, and Edited Nearest Neighbor (ENN).

**Table 4**: Performance metrics of the logistic regression classifier using several resampling methods. Acronyms used are defined below

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | RUS | | RUS/T-Links | | ENN (k=3) | | ENN (k=10) | | Oversampling SMOTE | |
|  | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| PP | 0.501 | 0.493 | 0.5 | 0.497 | 0.004 | 0.004 | 0.004 | 0.005 | 0.5 | 0.004 |
| NP | 0.499 | 0.507 | 0.5 | 0.503 | 0.996 | 0.996 | 0.996 | 0.995 | 0.5 | 0.996 |
| SS | 12295 | 1367 | 12295 | 1367 | 1403623 | 155959 | 1385432 | 153937 | 2813274 | 156977 |
| Accur. | 0.836 | 0.816 | 0.832 | 0.83 | 0.997 | 0.997 | 0.997 | 0.997 | 0.833 | 0.913 |
| Prec. | 0.906 | 0.884 | 0.896 | 0.891 | 0.828 | 0.829 | 0.87 | 0.869 | 0.897 | 0.036 |
| Recall | 0.75 | 0.723 | 0.751 | 0.75 | 0.322 | 0.309 | 0.351 | 0.372 | 0.752 | 0.728 |
| TNR | 0.922 | 0.908 | 0.912 | 0.91 | 1.0 | 1.0 | 1.0 | 1.0 | 0.914 | 0.914 |

**\*PP = Positive Proportion; NP = Negative Proportion; SS = Support Size; Accur. = Accuracy; Prec. = Precision; RUS = Random -Sampling; T-Links = Tomek Links; ENN = Edited Nearest Neighbor; SMOTE = Synthetic Minority Over-Sampling Technique**

In random under-sampling (RUS), some records of the majority class are randomly removed creating a smaller, balanced dataset. The smaller dataset also means a reduced run-time cost, especially significant when dealing with big data. The drawback of this method is there is information loss with the removal of data from the majority class. Applying the RUS to the mortality/non-mortality dataset increased the performance of the logistic classifier as indicated by performance metrics. The precision and recall increased from 0.76 and 0.32 in the imbalanced dataset to 0.88 and 0.72 respectively.

TomekLinks provides a guided undersampling/oversampling of a dataset by identifying and removing “noisy” data (TomekLinks) at the boundary of classes. Elhassan et al. (2016) suggest removing noise observation from majority class followed by RUS improve the performance of classification by reducing the chance of information loss. The precision and recall increased from 0.76 and 0.32 in the imbalanced dataset to 0.89 and 0.75 respectively. The performance showed a slight improvement from randomly under-sampled dataset only. A more significant improvement using TomekLinks removal would have been possible if there were more TomekLinks in the original dataset.

In Edited Nearest Neighbor (ENN), the majority class is under-sampled by removing points whose class label differs from a majority of its *k* nearest neighbors. Using *k* = 3, the number of data points decreased slightly from 1562931 to 1552751 with no improvement to the imbalanced data.

The performance metrics show that RUS/TomekLinks gives better results. The confusion matrix, precision-recall curve, and ROC for dataset balanced using RUS/TomekLinks are shown in Figure 9.

|  |  |  |
| --- | --- | --- |
|  | **Train Set** | **Test Set** |
| **Confusion Matrix** |  |  |
| **Precision-Recall Curve** |  |  |
| **ROC** |  |  |

**Figure 9**: Confusion matrix, precision-recall curve, and ROC for the train set (left column) and test set (right column) indicate good performance results of the classifier trained on the balanced data (using TomekLinks and RUS).

**5.3.2 Oversampling Methods**

Oversampling methods employ algorithms to increase the number of data points in the minority class to balance the minority/majority proportion. The two major drawbacks of oversampling are the increased possibility of overfitting and increased learning time due to increase in learning examples.

In this project, Synthetic Minority Oversampling Technique (SMOTE, Chawla et. al., 2002), a method that oversamples the minority class by creating synthetic examples, rather than the traditional way of oversampling with replacement. The data can be split into training and test sets before or after applying SMOTE. In this project both methods are attempted for comparison. If SMOTE is applied after splitting the data, there is a drawback that, the test is performed on synthetic data. The precision and recall for the SMOTE model, after splitting the data, are 90% and 75% respectively. If SMOTE is applied before the splitting of the data, the model is affected by the data imbalance between the positive and negative classes. The performance metrics for the oversampled dataset in Table 4 show the accuracy (91%), recall (72%), and TNR (91%) are good. However, the precision (3.6%) is very low, which is indicative of the high number of false positives (the true positive and false positive for the test set are 497 and 13,483 respectively). The SMOTE algorithms approach, compared to traditional oversampling by replacement, is to create a larger decision region. This could be the reason for the increased number of false positives. If the interest were only on recall (correctly predicting most of the infants who will die) without worrying about the cost of low precision (incorrectly predicting many more infants will die), this model would be acceptable.

|  |  |  |
| --- | --- | --- |
|  | **Train Set** | **Test Set** |
| **Confusion Matrix** |  |  |
| **Precision-Recall Curve** |  |  |
| **ROC** |  |  |

**Figure 10**: Confusion matrix, precision-recall curve, and ROC for the train set (left column) and test set (right column (using SMOTE).

**5.4. Feature Importance**

The model built by combining TomekLinks and RUS is selected as the best model based on performance metrics. Features importance analysis (Table 5) applied to the model shows that admission to Intensive Care Unit, Cyanotic Congenital Heart Disease, Spina Bifida/ Meningocele are the three major attributes with high positive coefficients and hence high effect on infant mortality (positive class). APGAR score has a high negative coefficient showing that it has high effect size on infant non-mortality (negative class).

**Table 5**: Feature importance table showing the significance of attributes.

|  |  |  |
| --- | --- | --- |
| Attribute | Coefficient | Absolute Value of Coefficient |
| AB\_NICU | 1.27 | 1.27 |
| CA\_CCHD | 0.75 | 0.75 |
| UCA\_SPINA | 0.63 | 0.63 |
| UCA\_DOWNS | 0.52 | 0.52 |
| UCA\_OMPHA | 0.51 | 0.51 |
| APGAR5 | -0.50 | 0.50 |
| UCA\_ANEN | 0.46 | 0.46 |
| MBRACE | 0.44 | 0.44 |
| CIG\_REC | 0.42 | 0.42 |
| UCA\_CELFTLP | 0.38 | 0.38 |

**6. Conclusions**

In this project, a logistic regression model is built to predict infant mortality/non-mortality based on routinely collected data during pregnancy, and delivery of a child. Due to the inherent imbalanced nature of the dataset, various under-sampling and over-sampling methods were explored to come up with the best resampling based balanced data.

Performance metrics indicate that pre-processing the data by removing TomekLinks followed by random undersampling gave the best results. Oversampling of the data (SMOTE) gives good results in accuracy, recall, and TNR, but the precision is very low. The SMOTE based model can be used if the main objective of a project is to correctly identify most infants that die without worrying about many more infants that will incorrectly be classified to the mortality class.

**7. Recommendations**

Hospitals and clinic collect huge amount of data related to pregnancy and delivery. These troves of data can be used to predict infant mortality using machine learning methods. The effectiveness of such methods depends on the availability and accuracy of data. Parents, policy makers and personnel assigned to collecting data should be diligent in collecting data and constantly analyze its significance to help understand and mitigate infant death.

The results from the project indicate that infant mortality can be predicted with high accuracy. The precision for the best model in this project is 89%. This means the prediction from the model is correct 89% of the time. The recall for the best model from this project is 75%. This means the model would correctly predict 75% of infant death. The accuracy of the model is 83% for both negative and positive classes, which means the model’s prediction for an infant mortality/immortality is 83% accurate. Responsible medical personnel can use the model or an improvement of the model to predict infant death.

One of the many measurements taken after child delivery is APGAR score. The data have shown that there are mortality cases for infants with high APGAR score. The major causes of deaths for infants with high APGAR score are sudden infant death syndrome, anomalous causes of death and sudden suffocation. A special machine-learning project is recommended to understand the first two causes of death. An intensive education is recommended to new parents of the dangers of suffocation and how to avoid it.

**Future Work**:

1. Expand the dataset to years other than 2008. This will show changes in the dataset with time as well as any change in the actual health data.
2. Apply other machine learning algorithms such as Support Vector Machines (SVM) and tree-based algorithms to compare with the logistic regression.
3. It would be interesting to explore whether there is a connection between mortality and other factors such as geographical location and demographics.
4. Use other resampling methods such as ADASYN.

**References:**

Aristaran, M., and M. Tagas (2013): Introducing Tabula, <https://source.opennews.org/articles/introducing-tabula/>. Accessed Dec5, 2017.

Chawla, N. V., Bowyer K. W., Hall L. O., and W. P. Kegelmeyer (2002): SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, 321-357.

Elhassan T, Aljurf M, Al-Mohanna F, and Shoukri M (2016): Classification of Imbalance Data Using Tomek Link (T-Link) Combined with Random Under-Sampling (RUS) as a Data Reduction Method, Journal of Informatics and Data Mining.

Lemaitre, G., Nogueira, F., and C. K. Aridas (2017): Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, Journal of Machine Learning Research, Volume 18, pp. 1 -5.