Infant Mortality Prediction

Using Pakistan Demographic and Health Survey (2017-2018)

Muhammad Uzair Aslam

5/8/20

Foundations to Data Science

Table of Contents

Introduction

The Infant Mortality Rate (IMR) is the number of infants per 1000 that do not survive until their first birthday. It is an important metric providing information not only about infant health, but also measures the society's general health status. Protecting the live of newborns has been an issue in public health, social policy and development economics.

Problem Statement:
To predict the infant mortality from Pakistan Demographic and Household Survey (2017,2018).

Literature Review

The works in this field includes important contributions (Saravanou, Clemens , Nicholas , Dolores , & Dimitrios , 2019).The research focuses on solving on binary classification task of infant mortality using birth certificate data and aims to establish most critical features of infant deaths. The best results were given with Boosted trees, highest AUC = 0.85 and recall = 0.77. Another important work is (Santos & Deborah Ribeiro Carvalho, 2018). This work aims to evaluate predictive methods for infant mortality in the State of Paraná between the years of 1996 and 2014. MLP model described in this work gave result with 97.7% accuracy which was the best with an average of 2.5% error between years of 2011 and 2014.Another important contribution is (Ghahfarokhi, Jamil Sadeghifar, & Mosayeb Mozafari, 2018) .They aim to use data mining technique in identifying accurate predictors of (low birth weight) LBW. The results of ANOVA showed that neonatal weight was higher among mothers with weight range of 84-110 Kg. The random forest algorithm showed that gestational age less than 36 weeks was the main predictor and number of foetuses, preeclampsia, and premature rupture of membrane, placenta previa, the number of pregnancies and the degree of mother education were other predictors of low birth weight. Several other researches focus on building classification models for predictive analysis with Infants Health Cards. The results show that APGAR score is one of the important medical risk factors contributing to infant survival. My proposed model unlocks a new dimension in this field. I have used the Pakistan Demographic and Household Survey 2017-2018 to build the classification models. This was the first-time survey data was used to model this problem and different socio-economic factor were significant which were previously not worked on has come to light. There has been significant relation in an infant "currently breastfeeding" to the mortality rates. My results outperform the previous models developed in this field with Light Gradient Boosting Machines gave result of 99.8% accuracy with recall 100% and AUC = 97.9%.

Datasets

The data collected is from Demographic Health Survey (DHS) program organized by USAID. The DHS program is recognized internationally for collecting and disseminating precise nationally representative data on fertility, maternal and child health and nutrition. I took a permission from DHS and wrote them the purpose of my project on these data sets. I thank them for giving me authorization to use datasets. My data set includes both rural and urban population of Pakistan. The survey followed a stratified two stage sample design where first stage involved selecting sample points(clusters) consisting of enumeration blocks while second stage involved systematic sampling of household. The survey was carried out in 561 clusters. The data contains approximately 50,500 sample points with 1186 features. Each survey sample includes features on fertility, family planning, maternal and child health, gender, HIV/AIDS, malaria, nutrition and other socio-economic factors. The class variable for predicting mortality presented in data is: **"**Child is alive**".** There were two datasets used in the predictive analysis. First was Pakistan: Standard DHS, 2017-18 for training the model and it was tested on Nepal: Standard DHS, 2016.

Methodology

I am interested in predicting infant mortality using the survey data records. I use a set of observation Xi (the variables from health survey) to predict the outcome Yi (whether the infant in survival class or not) using classification models where:

Yi = {0, if infant "i" belongs to Survival class, 1, otherwise

I trained different models in the two classes (Survival – Not Survival). The Pakistan: Standard DHS, 2017-18 was divided in two parts. I used a threshold of (80-20) where 80% of the data was used to train models and 20% of the data was for testing of those models. I used (i) **GNB**: Gaussian Naive Bayes, (ii) **SVM**: Support Vector Machine, (Scholkopf, Bernhard; Robert Williamson; Alex Smola; John Shawe-Taylor; John Platt) (iii) **RF**: Random Forest Classifier (iv) **NN**: a 3-fully-connected-layers neural network,(v)**LGBM**: Light Gradient Boosting Machine and (vi) **LR**: Logistic Regression (more details in Table 1).

In this study, I planned to build different models and explore relationship between them. Furthermore, different socio-economic factors are also examined which were significant in the classification task.

Experimental set up

Data uploading and extracting infants sample points

The data received for Pakistan: Standard DHS, 2017-18 from the DHS program was from the category of Births Recode**.** This category included data sets which were in line with my problem. The name of the folder was PKBR71SV.ZIP which included file named PKKR71FL.SAV. The data was received in the SPSS file format which was later transferred in Python using package pyreadstat 0.3.3**.** Initially, the data had 50,500 sample points of the whole survey. But given my problem, I extracted those sample points which were infants. After extracting the infants sample points the data was reduced to 2486 data points with 1186 features. The class value "Child is alive" was coded with B5 in the data. It was a highly imbalanced distribution (1:19) ratio of child not surviving to surviving. The total number of Survival infants were 2360 samples and non-survival infants were 126 samples only. This makes the data highly imbalanced. There were no null values in predicting class hence a separate file named Target.csv which only had one feature: Child is alive was made. Apart from this, another file was made to keep the track of the remaining 1185 features. This file only included the descriptions of 1185 coded variables of the data which was named name_labels.xlsx. The whole data without the predicting class was stored in infants_first_copy.csv.

Preprocessing

The next step was to analyze and clean the data. A great amount of time and energy was used to make sure that the data was cleaned without losing the significant information. Among the 1,185 features only 97 features were numeric with some important features such as Women's individual sample weight. This meant significant data cleaning and wrangling was required to transform the remaining object features into numeric datatypes. There were around 104,000 null values in the data which had to be imputed. Since there were only 2,486 sample points, so the none of the sample points were dropped and all the values were imputed. Since analyzing the 1185 features at a time was cumbersome, I analyzed 100 features at a time and following criteria was used to drop irrelevant feature.

- Features that are random generated numbers like CaseID, Country code and Phase etc
- Features that had more than 60 % of empty values
- Features that leaks information from future such as (age of death, reason of death etc)
- Features that had redundant information.
- Features which require extensive feature engineering such as date columns etc.

After using the above critrerea,866 features were dropped successfully from the data. This was done without losing any sample point and thus maintained the statistical information in the data. After removing the irrelevant features, I saved the file as infants_updated_1.csv. This file had 320 features apart from the predicting variable.

Before moving to imputations, I wanted to make sure that my data types were correctly defined. There were 282 features which had datatype as object and remaining 38 variables had datatypes as Float64.

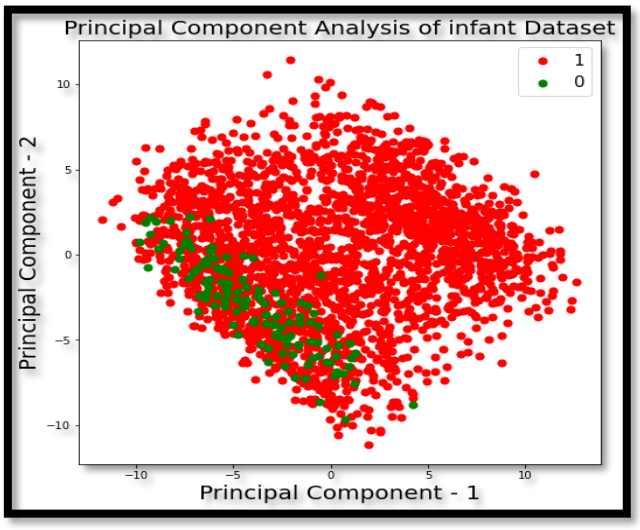
Starting from the object datatype, feature cleaning was performed to make their data type as Float64.This required significant feature cleaning and data wrangling. In this process all the data values termed as "not a dejure resident" were treated as null values as this was reported by survey specialists. Furthermore, values such as "flagged cases" and "don't know" were also treated as null values. After successfully converting to numeric attributes, the next task was to convert the remaining data types from Object data type to Boolean data type. As significant attributes were in this category, after converting to Boolean data type the remaining attributes were of categorical ordinal datatype. Finally, the remaining categorical nominal attributes were dummy encoded. In this process, I dropped the first attribute to avoid multi-collinearity. After applying dummy encoding the features increased from 320 to 489 features, this means that 169 more features were added after dummy encoding.

After making the required adjustments in datatypes, the next step was to impute the missing values. There were around 104,000 null values in the data which had to be imputed. Since there were only 2,486 sample points, so none of the sample points were dropped and all the values were imputed. I started imputing first with the categorical ordinal attributes. The categorical values were imputed using Sci-kit learn Simple Imputer with mode. Next I used Sci-kit learn Iterative Imputer: with maximum iterations =10, nearest features = 4 and initial strategy = "median". These parameters were set keeping in mind that the data was highly skewed in nature. Finally, after cleaning my data with correction of data types, null values, feature cleaning and wrangling, the file was saved to infants_updated_2.csv with 2486 sample points and 489 features.

Dimensionality Reduction

Before applying any machine learning model, next step was to reduce the dimensionality of the data. Therefore, **PCA**: Principal Component Analysis was performed. Before applying PCA, normalization of the whole data was performed using the Sci-kit learn Standard Scaler. First PCA was performed with components = 2 (figure 1)

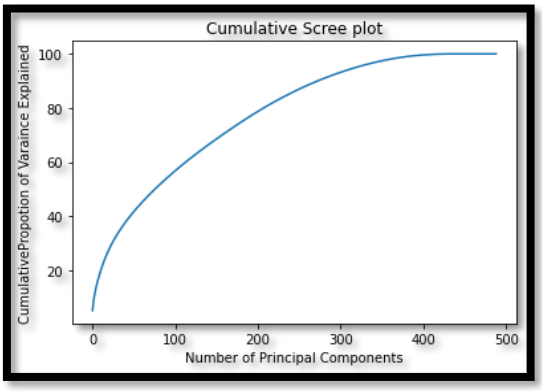
Figure 1: Visualization of Two Principal Components



The first two components show the spread of the data with 1(Red) indicating Infants that belong to the survival class and 0 (Green) indicating Infants did not belong to the survival

class. Their explained ratios were 0.0513 and 0.0302. This meant that two components had a combined proportion of 8% variance explained. (figure 2)

Figure 2: Visualization of Cumulative Scree Plot.

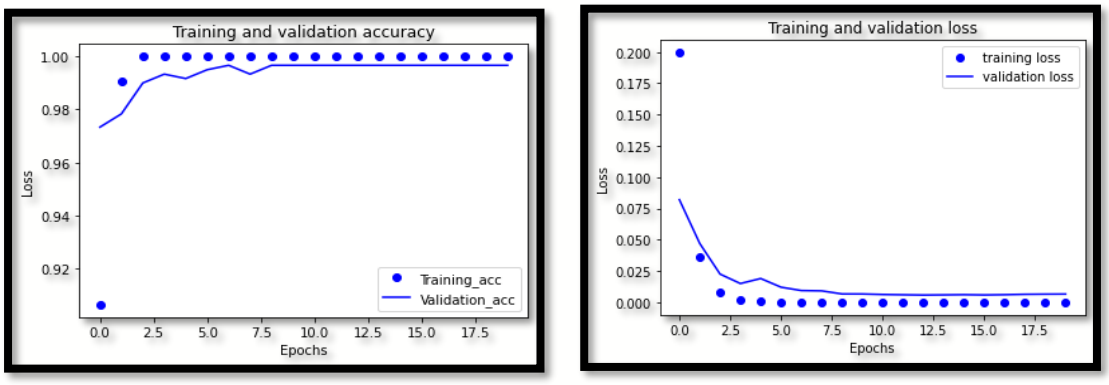


Looking at the plots gives a comprehensive picture of the number of Principal Components and proportion of variance explained. Therefore, I decided to keep components =321 as it retained 95% of the variance. Finally, after applying PCA the data was saved to infants_updated_4.csv

Training

As discussed earlier the data was divided into training and testing. In training data 50% of the data set was used for training where 30% was reserved for validation. The model was tested on 20 % of the data set. I used a combination of approaches to produce state of the art results. Hence the I trained **LR**: Logistic Regression and on the data reduced by PCA. **RF:** Random Forest Classifier, **LGBM**: Light Gradient Boosting Machine, **NN**: Neural Networks, **GNB**: Gaussian Naive Bayes and **SVM**: Support Vector Machine were trained on the original data set. This was done to maximize the Accuracy and ROC AUC scores. (See figure 3 For training and validation accuracy of Neural Network).

Figure 3: Training/Validation accuracy and loss of Neural Network with number of epochs

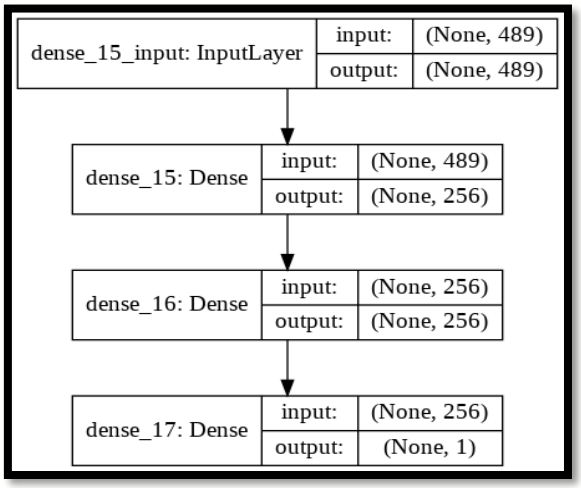


Parameters setting

This step was one of the crucial steps to produce best results. Hence hyper parameters were tuned using the Grid Search from Sci-kit learn. Furthermore, Pipeline class was also used to chain different operations. The models were trained using cross validation = 5. In topology of the i) **NN**: A dense 2 hidden layers neural network with the hidden layers uses rectified linear

activation function, the output layer uses the sigmoid activation function, the model is optimized using the binary cross entropy loss function, the rmsprop version of gradient descent with number of epochs = 20 and batch size =100 (See fig 4 for complete network).

Figure 4: Defining the Network Topology



For ii) **SVM**: one-class SVMs with RBF kernel with C = 10 and gamma = 0.001,iii) **LGBM**: n_estimators = 1000, num_leaves =10,max depth =5 with a learning rate =0.01, iv) **RF**: n_estimators = 100, bootstrap and oob_score = True, v) **LR**: solver = "lbfgs" and v)**GNB**: priors=None, var_smoothing = 1e-09.

Results and Discussion

Table 1:Evaluation for the classification models using 20% of testing data using all the features. Precision and Recall values refer only to the minority class.

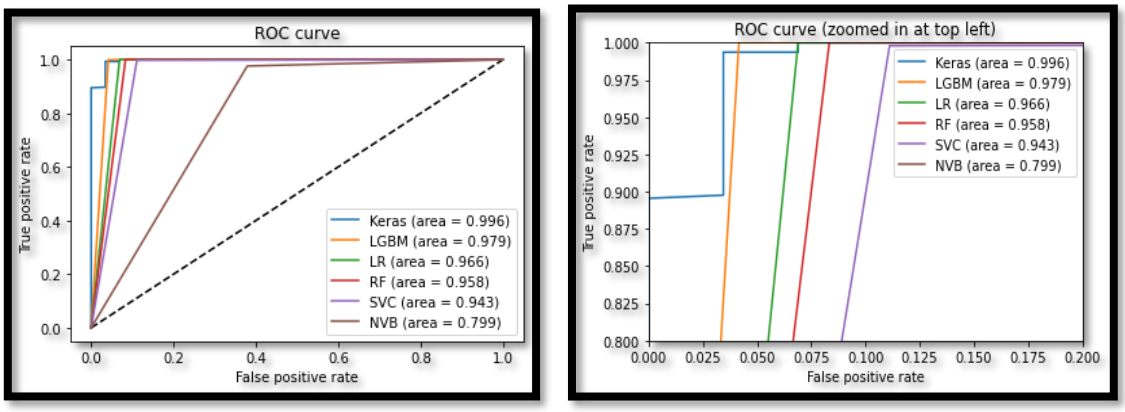
Metric	Neural Networks	Logistic Regression (reduced by PCA)	Support Vector Machine	Random Forest	Light Gradient Boosting Machine	Naïve Bayes
AUC	**0.996**	0.976	0.943	0.958	0.979	0.799
Precision	**1.000**	0.954	0.941	**1.000**	**1.000**	0.621
Recall	0.862	0.954	0.889	0.917	**0.958**	0.621
F1	0.926	0.954	0.914	0.957	**0.979**	0.621
Accuracy	0.992	0.996	0.994	0.996	**0.998**	0.956

Binary Classification Task

For this task, I applied different combinations of models with different feature combinations (All,10 best). Table 1 shows all best performing models using all features while Table 2 shows models trained on ten best features. The ten best features were selected using Simple Sequential Forward Selection (SSFS) from package Mlxtend. In such a task, I aimed for high recall, as my goal is not to miss any infant in high risk of mortality.LGBM outperform all the

models with recall = 0.958 and precision = 1.000.Furthermore,LGBM also scored high in F1 score = 0.979 and on accuracy = 0.998.Looking at other model's performances, Neural Networks and Random Forest got precision score = 1.000 which meant they were more conservative in predicting in the not survival class and were giving good result on predicting Survival class. Neural Networks produced highest AUC = 0.996 while LGBM giving second best AUC = 0.979. Logistic Regression produced balanced results on the Survival and Not Survival Class with AUC = 0.976. Random Forest and Support Vector Machines both gave par results, but Random Forest was better in terms of high AUC = 0.958 and Recall = 0.917. Lastly Naïve Bayes performed worst of all the classifiers with lowest scores.

Figure 4:ROC Curves with full view(left) and magnified from the top left(right).



Seeing the performance of LGBM on full features, I decided to use it on reduced features. The results in Table 2 shows that using only the 10 best features, it gave the same results as it gave on the full features. I tried to compare the results with Random Forest using 10 best features of Random Forest. Although Random Forest gave accuracy = 0.994 but from its AUC = 0.50, Precision and Recall scores it is a week classifier.

Table 2: Evaluation for the classification models using 20% of testing data using only 10 best features. Precision and Recall values refer to the minority class.

Metric	Random Forest	Light Gradient Boosting Machine
AUC	0.500	**0.979**
Precision	0	**1.000**
Recall	0	**0.958**
F1	0	**0.979**
Accuracy	0.994	**0.998**

Overall, these results show that LGBM surpassed all the classifiers including the Neural Networks and it can be considered as benchmark algorithm for this task.

Findings for feature importance

In general, the importance of feature indicates how informative and useful the feature was for building the classifier. Below table summarizes the top 10 features of Random Forest and LGBM.

Table 3: Feature importance by LGBM and Random Forest

LGBM feature importance	Random Forest feature importance
Women's individual sample weight (6 decimals)	Currently breastfeeding
Currently breastfeeding	Child's weight in kilograms (1 decimal)
Child's weight in kilograms (1 decimal)	Months of breastfeeding
Wealth index factor score combined (5 decimals)	Living children + current pregnancy
Wealth index factor score for urban/rural (5 decimal place)	Sons who have died
Received POLIO 1	Birth weight in kilograms (3 decimals) of Child
Marriage to first birth interval (months)	Duration of breastfeeding
All woman factor - wealth index	Received POLIO 0
Preceding birth interval (months)	Child's age in months
Child's height in centimeters (1 decimal)	Daughters who have died

Both classifiers consider “Currently breastfeeding” important feature in predicting Mortality. Furthermore, “Child’s weight in kg” is also of high importance in the predictive analysis. Lastly 5 features in the best 10 are highly related to mother and household incomes which shows the contribution of socio-economic factors in infant mortality.

Testing and Deployment

After training and analyzing, my goal was to make my model as generalized as possible. I used Nepal: Standard DHS, 2016 to test my model on completely unseen data. The data included 378 sample points. After preprocessing, I fit my trained Random Forest algorithm. The model was able to classify all the sample points correctly and gave 100 percent accuracy. This motivated me to deploy the model on web. Finally, after making Flask API, I successfully deployed the model on Heroku for public use.

Conclusion and Future work

In this study I tried to explore and analyze the infant mortality problem. It is a very challenging classification problem due to high imbalance in the data(1:19).This is a novel idea because previous works have used health cards to predict infant mortality but this is the first time that classification models were built on demographic and health survey data sets. There is huge potential in this research area as it combines different fields like medical, health, policy making. This makes the task challenging yet critical for the developing nations such as Pakistan where mortality rates are high. Future work in this field focus on extending

my study to more general features such as maternal factors. Moreover, household distribution of income can be connected to the infant mortality predictions. Furthermore, infant mortality models are to be linked with Poverty which will open a whole new paradigm for economic development and data science.

References

Ghahfarokhi, S. G., Jamil Sadeghifar, & Mosayeb Mozafari. (2018). A model to predict low birth weight infants and affecting factors using data mining. *J Bas Res Med Sci* , 5(3):1-8.

Santos, A. B., & Deborah Ribeiro Carvalho. (2018). PREDICTIVE MODELS FOR INFANT MORTALITY IN THE STATE OF PARANÁ. *Iberoamerican Journal of Applied Computing* .

Saravanou, A., Clemens , N., Nicholas , H., Dolores , A.-G., & Dimitrios , G. (2019). Infant Mortality Prediction using Birth Certificate Data.

Scholkopf, Bernhard; Robert Williamson; Alex Smola; John Shawe-Taylor; John Platt. (n.d.). Support Vector Method for Novelty Detection . *Microsoft Research Ltd.*

Model Link: https://infantmortalityapi.herokuapp.com/