|  |  |
| --- | --- |
| Helping in the journey to a healthy Motherhood  Healthy Mom + Baby Datapalooza | Abstract  A step to reduce Indiana’s infant mortality rate and achieve the lowest rate among Midwestern states by providing concrete insights and solutions through an R-Shiny app that can be implemented to improve infant and maternal health in Indiana.  Team: We r family  Anjali Hegde (hegde@purdue.edu)  Snehalatha doddigarla (sdoddiga@purdue.edu)  srishti bisen (sbisen@purdue.edu)  sukhsagar jaiswal (jaiswal@purdue.edu)  vikas das (das104@purdue.edu) |

## Business Problem

Infant mortality is the leading indicator of the quality of any public health system. At 7.2 IMR, Indiana ranks 7th highest in the United States in terms of infant mortality, higher than the national average of 5.9 infant deaths/1000 births. Therefore, Indiana government now looks forward to reducing Indiana’s infant mortality rate and achieve the lowest rate among Midwestern states by 2024.

Many organizations in Indiana hold a vast repository of information in this regard which can be analysed to find possible reasons and solutions for high infant mortality in Indiana.

**Stake holders:**

The stakeholders who would benefit from the solutions to this problem are women who plan to have a child or are in the early/later stages of pregnancy, new moms, infants, hospitals and OB/GYN care providers.

## Analytics Problem

One initiative to reduce Indiana’s infant mortality rate and achieve the lowest rate among Midwestern states by 2024, is the OB Navigator program, a collaboration between ISDH, FSSA and the Indiana Department of Child Services DCS which builds a network of services and support to create healthier outcomes for both moms and babies.

Our Analytics solution strives to discern the trend of mortality by county and ethnicity and predict possible outcomes of infant mortality through Indiana’s repository of information to deliver powerful insight which can be used by the OB Navigator program officials to develop solutions to optimally allocate resources to help mothers. One limitation of our approach is lack availability of real time data.

## Data

The data source referred in our endeavor is the official website of Indiana State Department of Health. The datasets give us access to information of year wise infant mortality rate for every county along with leading causes of infant deaths in these counties for exploratory data analysis. For statistical model building we referred to Indiana births and Infant deaths [3] dataset which had information on case basis of every mother and infant born to the said mother, their lifestyle habits and outcome of infant mortality.

The dataset had missing values and therefore pre-processing was done to rescale and clean the dataset and the same has been shared as part of DSS. This dataset was divided in to 70:30 ratio to train and test subsets.

**Methodology Selection**

In our data, the target variable we had to predict was infant mortality indicator, which was either Yes or No. This made it a classification problem where we need to predict the probability of an infant’s death based on parameter’s likes mother’s age, child’s gender, whether the mother smokes or not, etc. We followed the below three analytical approaches for classifying the response variable:

a. Logistic Regression: We employed different variations of logistic regression technique to classify the target variable which will fit the data space exactly into two

b. Tree: We employed different variations of tree, such as Decision Tree, Random Forrest, XGBoost to classify the target variable by bisecting the data space into smaller and smaller regions

c. Deep Learning: We employed deep learning model to predict for the response variable

R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing and is widely used among statisticians and data miners for developing statistical software and data analysis.

For our case the largest data had over 700,000 observations which did not consume a lot of memory space, however the data was imbalanced. R provides a plethora of libraries to solve for classification problems in a concise and interpretable manner also allowing the functionality to balance your data by sampling. As we had to predict for infant mortality, which is a classification problem, employing variations of logistic and decision tree algorithms are appropriate

## Model Building

We ran the below four models on H2o:

Generalized Linear Model Random Forrest XG Boost Deep Learning

Since this is a classification problem, we need to evaluate models on the AUC value which is the ratio of False Positive Rate to True Positive Rate. Below is the evaluation table for the models which were built on 80% of the entire data and validated on the remaining 20% test data:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | AUC | R Squared | RMSE | Misclassification error |
| GLM | 0.8132317 | 0.0165654 | 0.06226796 | 0.010730 |
| RF | 0.8099258 | 0.019609 | 0.06217153 | 0.009545 |
| XGB | 0.8172307 | 0.01348314 | 0.06236547 | 0.008843 |
| Deep Learning | 0.7405095 | NA | 0.5591417 | 0.018333 |

1. Once the model was trained on Train data and validated on Test data, it is fitted on the data which was generated by the user input by hitting ‘Click’ on different parameters which are used for predicting response.
2. Based on the evaluation table we have selected XGBoost as the preferred modelling algorithm as it has the highest AUC value and misclassification error across all other techniques.
3. Assumption: In our data, many columns had “Unknown” values which we imputed with mode of the column with respect to value of the response variable. For example, if the column Mother’s Marital Status is “Unknown” and the mode is Never Married for when Child Mortality Ind is Yes, then “Unknown” is replaced with Never Married.

## Functionality

The RShiny app has three tabs:

1. County Map:
   1. It has a dropdown for Year which ranges from 2011 through 2018
   2. With a selection of the Year, mapchart is generated for the state of Indiana which shows the infant mortality across counties in that respective year
2. Graphs:
   1. It shows the descriptive view of the data highlighting relationships between different variables with infant mortality
   2. The first graph shows infant mortality by Top 10 cause of death (COD) aggregated across years and county
   3. The next chart shows the number of OB-GYN providers across counties
   4. The last chart looks at Obesity across pregnant women across counties
3. Prediction Probability:
   1. It is an interactive tab which provides the infant mortality probability based on the mother’s profile
   2. The user can add inputs regarding the profile of the mother to understand the risk to the baby which includes information like n umber of babies, mother’s age, marital status, county type, if the mother smokes etc.
   3. The output shows the prediction which takes either 1 or 0 and the probability of infant mortality
4. Useful R packages which we used for building the RShiny tool are below:

Shiny Dplyr ggplot2 Shinythemes choroplethr choroplethmap

Readxl Caret data.table sqldf h2o Markdown ggthemes

Maps mapdata

1. Enhancements which we could have added to the RShiny app:
   1. Make the Indiana map chart interactive with drill down functionality when the user clicks on a county
   2. Create a tab, where the user can add data which could be then appended back with our original data and then the prediction is run
   3. User can create a profile and the download the data for future reference

## GUI Design and Functionality

1. The County Map tab of the app is designed to give an intuitive view of the state of Indiana and the Year drop down allows the user to check status of a County over the years
2. The Graphs tab is a descriptive summary for the OB Navigator to concentrate their resources on the appropriate counties
3. The Prediction Accuracy tab is a real time interactive dashboard where the OB Navigator can create multiple profiles and understand which variables are more important than others and the risk potential of the profiled mother and child
4. The app runs without any error

**Conclusion**

By understanding the underlying causes of infant mortality generated by our R-Shiny app and implementing suitable solutions to mitigate these causes, the state of Indiana can become a healthier and happier place to live and grow for mothers and their babies and can collectively realize its goal of becoming the state with the lowest rate of infant mortality among mid-Western states by year 2024.

**References:**

[1] <http://indiana.himsschapter.org/healthy-mom-baby-datapalooza-event>

[2] <https://hub.mph.in.gov/dataset/infant-mortality-by-county-year-and-underlying-cause-of-death>

[3] <https://hub.mph.in.gov/dataset/indiana-births-and-infant-deaths>

[4] <https://hub.mph.in.gov/organization/b72f3c69-1b03-474d-ab36-1ac2da7d91bb?tags=Datapalooza>