**Report: Reduction of the Under-Five Mortality Rate among the Empowered Action Group**

Divya Alukal

Data Analytics Program, Franklin University

DATA 695: Capstone

Dr. Deanne Larson

March 18, 2023

**Business Problem**

In this report, I will describe the analytics problem that I undertook on behalf of the Ministry of Health and Family Welfare of the Republic of India. The task was to create a model that could use health and demographic factors investigated in the 2012-2013 iteration of the Annual Health Survey (AHS), a comprehensive survey implemented in nine Indian states with particularly poor health outcomes, to predict the ratio of children in each district who die between birth and their fifth birthday per 1,000 live births. This ratio is known as the Under-Five Mortality Rate, hereafter U5MR (UNICEF, 2023).

To generate the final predictive model, I had to explore the data and analyze them in various ways, after which I was ready to generate multiple models and evaluate them to determine which would the best one. Some of the challenges I faced in this endeavor were caused by the size of the dataset. The preponderance of features, or variables, led me to spend ample time in the exploratory data analysis phase selecting a subset containing just a fraction of the features that I would use to create models. On the other hand, the low sample size of fewer than 300 districts was not ideal, as having more data points would have contributed useful information to the training and testing of the model. Nonetheless, I made my best efforts to create a predictive model that would identify the factors most relevant to this grave issue.

**Ethics and Security**

Because the data collected through the survey were aggregated at the district level in the data set, there are no security or ethical issues anticipated. Many variables included in the data set could potentially be sensitive if associated with an individual person; a few examples of these are marriages among females or males below the legal age, persons suffering from certain illnesses, and women with two children wanting no more children. However, data for all variables except state and district were provided as numbers or rates, and many residents of each district were surveyed. Therefore, there is no chance of a particular response or trait being attributed to any individual or household.

**Data Exploration and Cleaning**

The data provided to me had to undergo preliminary examination and changes in form before they were ready to be used for modelling. After a visual inspection of the dataset, I checked the dataset for empty cells and null values by feature. The reason for this is every sample in the set must have a non-null value for every variable to effectively do data analysis and to run Python, the programming language through which I completed the data analysis. Neither a row pertaining to a sample nor a column pertaining to a variable in the dataset can be partially filled. However, the question arises of what to do when null values are present. In general, one can either impute missing values – assign a value to them based on values for the feature of other samples – or remove from the dataset entirely a sample or feature that contains null values. Imputation is useful in that it allows us to retain data rather than eliminating it and is typically based on related values; for example, the column average or the median of the existing values above and below a particular missing value may be assigned in order to make the null value non-null. The downside of imputation is that it falsifies the data, although the falsification is justified. Since imputation is based on other values, the values it provides tend toward averages and are by nature not real. Nonetheless, imputation is generally preferable to removal of samples or features, a step usually reserved for when a row or column contains many null values and is deemed a poor provider of information. In the case of this dataset, which already has a low number of samples, it would be highly undesirable to remove any samples; although I did plan to later eliminate variables, I hoped to do it on the basis of lack of usefulness to the analysis and not lack of values. Fortunately, neither was required. No district’s data were missing a very large number of values and, although close to one-third of features had at least one null value, only a miniscule portion had at least 20% of their data missing. Ultimately, I chose to impute all missing values in the dataset based on the average for the feature of each sample’s respective state, with the assumption in mind that, despite variations in healthcare and other factors within each state, there would be some statewide commonalities that would make each state’s values cluster together rather than being on par with other states.

Following this came one of the most time-consuming and riskiest steps in the entire project, which was to use data visualization and feature engineering to select a subset of the over 600 variables in the original dataset to use in the analysis. Aside from potential computing problems from running Python models that utilize hundreds of features at once, it would not be very useful to do so. It would be difficult to order, assess and visualize the relative importance of each feature to the model and, moreover, would not create sustainable models. All the data that were collected through the very comprehensive AHS during the three years in which it was administered might not be available in the future; for example, information such as current usage of condoms or number of persons suffering from Acute Respiratory Infection would likely be more difficult to consistently collect than factors such as literacy rate or children engaged in work. There would be no point in creating a predictive model based on data that may not exist in the future. Another difficulty with the dataset was that many of the features were split into three or nine variants to account for differences by gender and/or location. Although having this plenitude of data was to be appreciated, using multiple variants of most of the features would overwhelm the analysis and would likely not contribute enough information to the model to be justified. Therefore, I had to determine for which of the features I should use split variants rather than just the basic feature.

As mentioned, this process was risky because it would rule out most of the original hundreds of variables from the final dataset that would be used to create the predictive models. Thus, I relied on my subject matter knowledge to determine which variables from the original dataset were most likely to be highly related to and predictive of U5MR and sustainably collectible. While all 643 features likely have some bearing on the target, I deemed something like the rate of mothers whose blood had been tested for hemoglobin to likely have less influence on U5MR than something like median age at first live birth of women, and thus suitable for elimination. In this vein, I went through every feature in the dataset and either added it to one of two tiered lists or removed it from consideration. The first tiered list was for variables which were most likely to be relevant to the model and the second was for those which might be relevant but were not as likely. Each list had dozens of variables for consideration and only contained basic forms of features without gender or location breakouts, since I presumed that distinctions among basic features would generally be greater than distinctions among variants of one feature. I then created correlation matrices between predictor variables and the target to visualize and quantify the level of correlation between them, making note of the predictors that had relatively high correlations (approximately 30% or more) with U5MR.

I ended up with 14 such predictors after using only a portion of the variables from the first tier that had minimal overlap in information. The four with the highest correlations to the target were, unsurprisingly, features related to death and mortality rates. They would be included in the subset of variables used in modelling, but do not contribute much to understanding the root causes of U5MR. For the remaining 10, I visualized the distributions of their gender and location variants to see if the differences were significant enough to warrant the variants also being included in the data subset. The final visualization was a chart of U5MR across the nine survey states. As the range of rates was over 40%, this was a clear sign that state was an important factor in U5MR and should be included in the new dataset used for modelling. At this point, the dataset had 26 features, none of which had had at least 20% of their values imputed.

After this, I did the final steps to prepare the dataset for analysis. Next, I standardized all values in the dataset. This is a process that brings all numerical values onto the same scale so that features measured on a larger scale, such as population, which is usually in the tens or hundreds of thousands, are not given more weight in a model than those measured on a smaller scale, like mean female age at marriage, which typically is in the teens or twenties, or a percentage like literacy rate, which necessarily has a value of 100 or less. I replaced the state variable, which identified each sample with a word, with dummy state variables that, for each of the states, provided a numerical yes-no indicator to show whether a sample belonged to the state or not. This was necessary because the predictive models could not be created using data based in words. A precaution taken while doing so was to remove the dummy variable for one of the nine states in order to prevent linear correlation between the predictor variables; for this reason, a dummy variable for the state of Assam is found neither in the final dataset nor in any portion of the model development and evaluation. In the end, there were 32 variables in the dataset. The last step of exploratory data analysis was to split the data set into two subsets: a larger training one which would be used to generate models, and a smaller test one which would be used to evaluate the models and ensure that they worked well for data that had not been used to create them. For both subsets, I separated the predictor values from the target values.

**Model Development**

After exploratory data analysis, I began model development. The aim was to create regression models that would predict a specific, accurate value for any given sample using patterns in the data established from the training data subset and then choose the best. I had seven model types in mind to try, although I knew that some of them may not be good predictors for these data. For each of these seven regressors, I used 10-fold cross-validation to fit the regressor to the training predictor data, that is, to make the regressor learn the patterns in the data and the type of target value that accompanied these patterns for each sample through a series of ten iterations. I then used the fitted models to predict target values for the test subset of samples. The purpose of this was to gain an estimate of which regressors would generally be best to take forward; however, this step would be repeated later once the models had been modified with better conditions. Based on the metric explained variance that measures the amount of variation in the target variable that can be attributed to predictor variables, I determined that the four model types with which to continue were Linear Regression (LR), Ridge Regression (RR) , Gradient Boosting Regressor (GBR) and Random Forest Regressor (RFR). Of these, the first two involve using all features to create a predictive line with as little distance from all data points as possible. The last two involve averaging the results of multiple decision trees, which function like flowcharts that redirect the prediction for a sample downward based on its values for various features. Although the four regressors differ in some ways, they all ultimately predict one continuous numerical value of the target variable for each sample.

The four model types to generate having been decided, they needed to be developed, trained and evaluated in depth. I used an exhaustive search to tune each of their hyperparameters, which are parameters that are set before the model starts to learn, that is, be trained with data. Changing hyperparameters from their default values can affect how well a model trains (DeepAI, n.d.). I then refit the models with their ideal hyperparameters to the training data. Using these versions of the models, I predicted target values for both the training and test subsets of predictor variables.

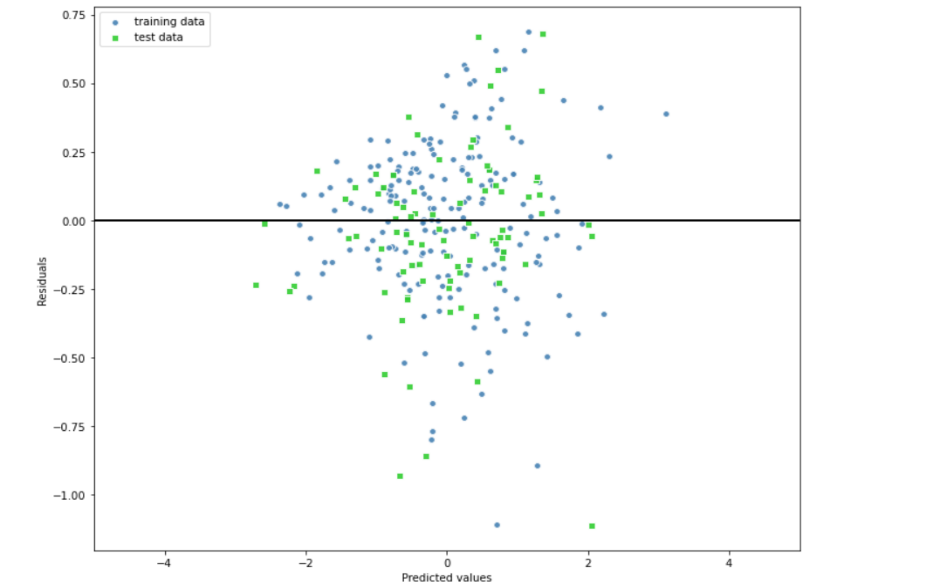
In addition to these four models, I used a process called Principal Component Analysis (PCA) to create another version of each model. PCA is useful for problems with many features because it reduces the number of features by encapsulating all of the information that they hold in new variables called principal components (Joliffe et al., 2016). In this case, the first six components encompassed approximately 80% of the information provided by the original 32 variables, so I limited the PCA versions of the models to using these six in training and testing. Although the obvious drawback of using PCA is that some information is lost, it adds some interpretability by reducing the number of features used to create the model.

**Model Evaluation**

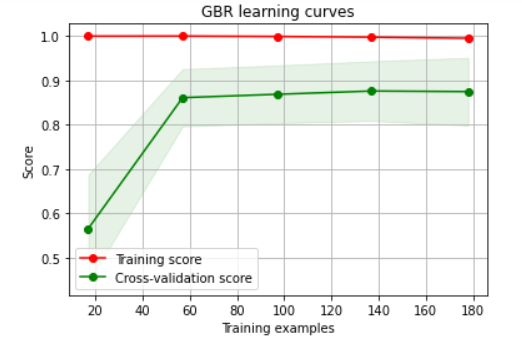
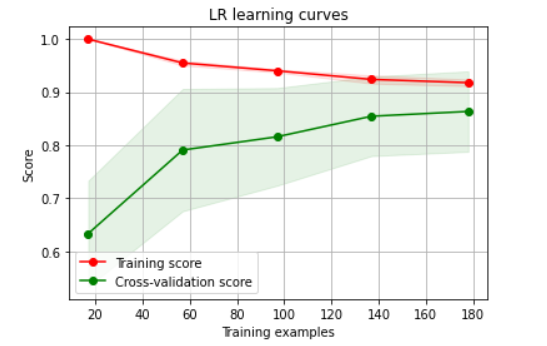
After creating each model, I evaluated it for accuracy and quality of information provided. The two primary metrics used for this were mean squared error (MSE), which measures the difference between predicted and actual values and penalizes larger errors, and coefficient of determination (R²), which indicates how much variation in the target variable can be explained by the predictor variables. To be sure that a model can generalize to data it was not trained on, it should perform similarly, or ideally better, on unseen data. Therefore, I looked for models having a lower MSE and a higher R² for the test data subset than for the training subset.

I also used various forms of data visualization to illustrate these metrics. For example, a residual plot is a type of scatterplot that shows the difference between predicted and actual values. There are two qualities desirable in a residual plot: concentration of points near the black horizontal line and lack of bias or patterns. This visualization is helpful in identifying if a linear model is inappropriate for the problem and how great the error in a model is. Another visualization used was learning curves, which show how the training and test scores based on explained variance change as the model is given more samples to use. In such a chart, the two lines will ideally approach each other as samples are added. This indicates that the model could be further improved by the collection of more data points, although in the case of this project, there are practical limitations to this. The learning curves for several of the models, however, showed training scores which barely, if at all, approached the test scores. Such instances aligned with other indicators that those models were not suitable for this problem.

*Residual plot for Linear Regression model. It has the second quality but not the first.*



*Linear Regression learning curve: promising. Gradient Boosting Regressor learning curve: discouraging.*



Other data visualizations used were specific to the model type. One of the key requirements of this project was to identify the factors most pertinent to U5MR so that they specifically could be addressed. For this reason, I used bar charts to show the relative importance of each feature to the output of the models. Interestingly, the results varied greatly by model; the RR model showed that, in addition to the four death- and mortality-related features, the features with the greatest effect on the model were a mixture related to birth rate, state, education and healthcare. The output of the GBR model, on the other hand, which performed very poorly overall, was close to 80% based on the Infant Mortality Rate; this information is unhelpful, as the causes of this feature are greatly intertwined with the causes of U5MR and would need to be understood on their own to lead to understanding of U5MR.

For each of the two RFR models, I visualized one of its 100 constituent trees for a depiction of the decisions that took place at each node of the tree. Those trees turned out to be too large to visually analyze in detail, but each led to dozens of values as potential results. The problem with tree-based regressors is that they cannot account for much distinction among feature values, as they flow downward through dichotomous splits. Unsurprisingly, the models based on the tree-based regressors RFR and GBR performed quite poorly.

**Conclusions and Recommendations**

The PCA RFR model was the worst of the eight created. Its MSE and R² values revealed that it generalized to unseen data extremely poorly and would not have been useful as a predictor of U5MR. The best model, based on its evaluation metrics and information provided about importance of features, was the RR model. The second-best model was LR. I found that there was 100% match between the predictions made by these two models for samples from the test data subset. For that reason, I combined the two into an ensemble regressor that would average the predictions made individually by each; in case of any discrepancies on new data, having two contributing models would strengthen the final prediction better than relying on one model. I recommend this ensemble regressor as the final predictor of U5MR in the nine AHS states. Additionally, the feature coefficients derived through the two contributing models indicate that literacy, crude birth rate, children’s school attendance and workforce participation, mothers’ prenatal care and children’s immunization are among the factors most highly related to U5MR. It is no secret that most of these have long been in dire need of improvement. Policies and programs should be implemented to increase general and health education for people of all ages, to reduce child labor, to aid citizens in family planning and to facilitate access to healthcare for children and their families.

As mentioned earlier, the low sample size was a disadvantage in this project. Learning curves for the non-PCA LR and RR models indicated that adding more training samples would benefit their performance. I do not recommend collecting data from other states other than the nine for whom this project was intended, since their socio-economic circumstances would differ too greatly from those of the original group to make their information useful in this context. Instead, I suggest conducting another round of the survey. More than ten years have passed since the original survey was completed in which political changes, societal shifts and a global pandemic, which all may impact U5MR, have taken place. Data from the second round could be used to create an updated predictive model and, depending on preliminary analysis, could also be used to augment the 2012-2013 dataset to create a more robust model with a double sample size.

**References**

DeepAI. (n.d.). *Hyperparameter*. DeepAI. Retrieved March 16, 2023, from https://deepai.org/machine-learning-glossary-and-terms/hyperparameter

Joliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, *374*(2065). <https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202>

UNICEF. (2023, January). *Under-five mortality*. https://data.unicef.org/topic/child-survival/under-five-mortality/