Predicting Maternal Mortality

Julia Eigner

5/6/2020

Contents

1. Introduction

Analyzing maternal mortality rates globally can indicate many areas for sub-analysis including a country's health system, income, education and urbanization. This project attempts to explore determinants of maternal mortality along these lines, using traditional covariates for health outcomes such as GDP and under-5 mortality as well as less traditional indexes such as the Global Peace Index and Human Developmen Index. General linear models, K-Nearest Neighbor, CART and Random Forest models are applied to explore data compiled from the World Bank, World Health Organization, Institude for Economics & Peace and the

UN Development Programme. The relative success of these models is compared using root mean squared error accuracy (RMSE) measures as well as r-squared. Feature-level analysis is conducted using partial dependecy plots as well as variable importance plots to explore predictor-outcome relationships as well as heterogeneity between countries.

2. Background

The goal of this project is to generate a model to predict country-level maternal mortality rates (measured per 100,000). More specifically, the goal is to explore the social determinants of maternal mortality to uncover heterogeneity among features that have strong priors. Specifically with public health research, there is a robust body of literature on the predictors of key outcomes such as maternal mortality. As such, this paper attempts to examine such priors for heterogeniety and interaction.

Importantly, maternal mortality rates indicate many factors besides a mothers' health status, as it is dtermined by socio-cultural factors of health including income, education and locality. Regional and sub-regional analysis of maternal mortality has indicated high variability in mortality based on race and income level. It is through this analysis that inequities in intra-country health outcomes as well as global inconsistencies can be found. Much of the existing literature utilizes proxy measures for health care utilization and quality for determinants of maternal health outcomes. Birmeta et al (2013) find that 87% of women in had at least one antenatal care visit during their last pregnancy in a study conducted in Ethiopia. Additionally, they find significant association between ANC attendance, used as a proxy for healthcare utilization, and socio-economic factors such as age at last birth, literacy and income, in line with the hypothesis for this model. In an analysis drawn from Demographic and Health Survey Data, community and individual level was used to identify determinants of the use of facility delivery care in Nigeria, non-facility delivery accounted for 14% of maternal deaths in 2014 (Ononokpono and Odimegwu, 2014). Region of residence, proportion of women who had a secondary education and several individual level factors

were significantly associated with facility delivery. Importantly, significant determinants of maternal mortality differ at the person, community and country level. Factors that predict sub-national mortality rates and statistically significant disparaties in intra-country analysis, such as density of health facilities, may not be a significant determinant in cross-country analysis (GBD 2015 Maternal Mortality Collaborators, 2016).

A regional analysis of under-5 mortality (U5M) in sub-Saharan Africa shows vast disparities among urban and rural populations. Holding economic growth and country-wide U5M constant "each percentage increase in urban population growth between 2005 and 2010 was associated with almost 170 more under-5 deaths per 1000 live births in rural areas compared to urban areas (95% CI: 66.33, 272.27, p=0.002)" (Beatriz et al, 2018). This finding alludes to urban growth as an important determinant of U5M at the individual level in rural areas. While this supports the notion that there are rural-urban disparities in health achievement, the study suggests that urban population growth is not associated with the probability of under-5 mortality at the country-level. This is affirmed in discussions below on the importance, or lack thereof, of the proportion of the population that is rurally based as a determinant for maternal mortality. Conversely, Kimani-Murage et al (2014) explored intra-country disparities in Kenya and found that the "urban advantage" no longer applied in Kenya, where child mortality rates in urban areas now exceed rural areas, likely due to high levels of informality and population density. This points to a possible interaction between income and rates of urban dwelling at the sub-national and individual level.

3. Data Sources and Cleaning

The outcome variable of interest, maternal mortality rates, were found at the World Health Organization Global Health Observatory Data Repository. Many of the features were found at the World Bank Open Data Repository which is the aggregating source for affiliate institutions such as UNESCO and ILO. Despite many sources avaialble for country level data for common indicators such as GDP and life expenctancy, the World Bank was preferred so

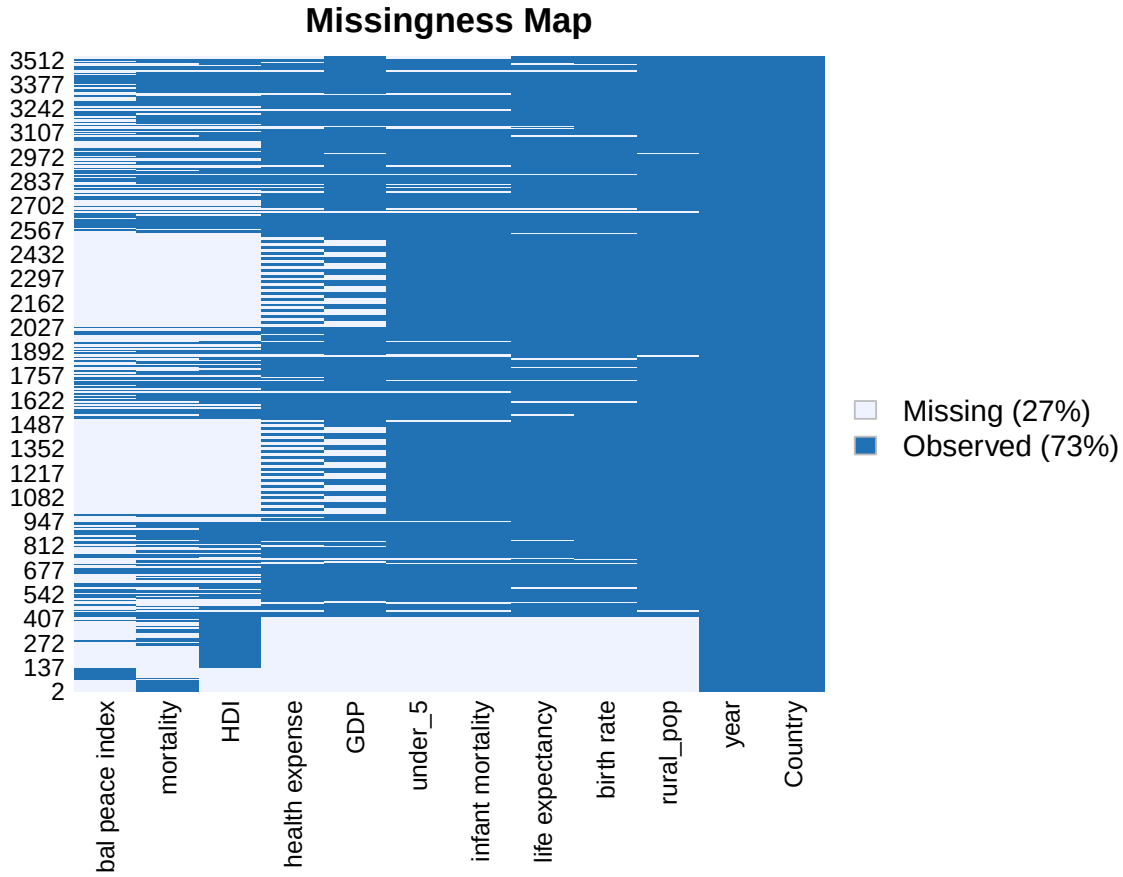
that naming conventions and data organization was as common as possible before cleaning began. Using a common datasource also increased the likelihood at the observation numbers for country-level data would be similar, preserving as much data as possible. Missingness, discussed below, was the cause of most of the data loss. The Human Development Index was found at the United Nations Development Programme website. Finally, the Global Peace Index was scraped from wikipedia, which references the Institute for Economics and Peace as the original data source.

Based on the availabe literature, education and use of health facilities were identified as key determinants of maternal mortality. Subsequently, HDI and health expenditure are considered key variables because they could serve as proxies for education attainment and health facility density. Whether these assumptions of suitable proxies hold is discussed in sections 5 & 6 below. A primary component of the Human Development Index is the "education index" comprising of expected years of schooling and mean years of schooling.

Variable	Description
mortality	Maternal mortality (per 100,000)
birth.rate	Crude birth rate (per 100,000)
GDP	Gross Domestic Product (current $US)
health.expense	Health expenditure per capita (current $US)
infant.mortality	Infant mortality rate (per 100,000 live births)
life.expectancy	Life expectancy at birth (years)
under.5	Mortality rate (per 100,000 live births)
HDI	Human Development Index
global.peace.index	Global Peace Index
rural_pop	Rural population (as a % of total population)

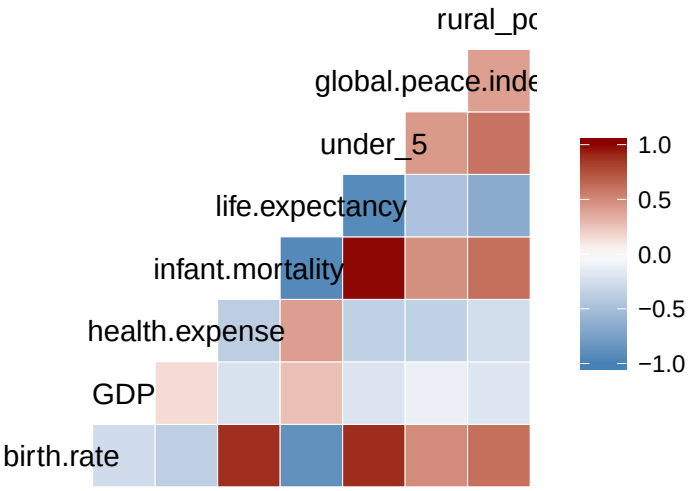
All data was imported at the country level, with years as features. Basic cleaning was conducted to match identifying variables, country and year, before a merge. Initial

cleaning involved stripping special characters "[" and whitespace from strings before mutating character variables to numeric. Once basic elements were cleaned, the datasets were pivoted long and joined. Once the data was long, the unit of analysis became country-year, with years 2010-2017 available for every country. The full dataset was analyzed for missingness and the potential for imputation. Originally the dataset contained 3,527 observations. 27% of the data was missing from the full dataset, as shown in the Missingness Map below. The majority of missingess comes from non-World Bank indicators: Global Peace Index, HDI and Maternal Mortality. Countries that were dropped from the dataset and excluded from subsequent analysis were limited to those that were missing for all years 2010-2017, meaning imputation would result in a value not based on means or neighbors from that country. After countries with data missing for all years were dropped, the dataset contained 1,080 observations which were used for analysis.



4. Analysis

Before beginning analysis the data was split into training (2010-2015) and testing (2016 & 2017) data. A recipe was constructed to impute missing values which, after missing countries were dropped, accounted for 6% of the dataset. The range was also normalized across the dataset. The distributions of the features in the training data and were normalized through pre-processing. Some correlational analysis is shown below and unsuprisingly birth rate, infant and under-5 mortality have strong, positive correlations with maternal mortality.



To prevent overfitting the k-fold cross validation method was used and the data was trained on 5 folds. Given that we are predicting a quantitative outcome rather than a binary outcome, the mean-square-error is used as the accuracy measure.

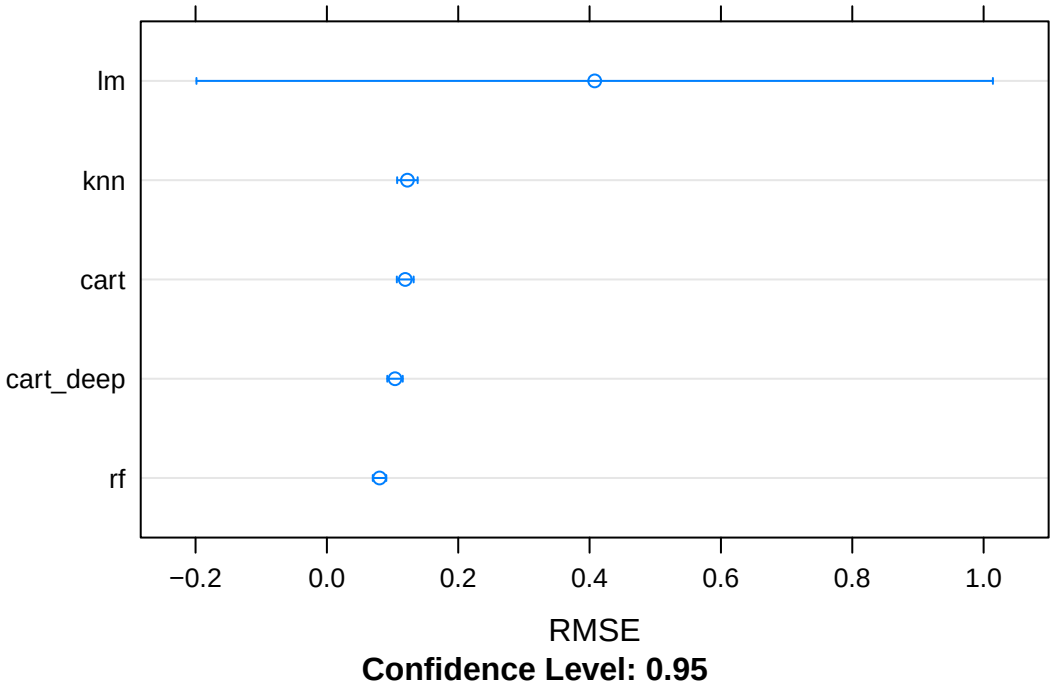
Four models were explored in this analysis: linear regression, K-Nearest Neighbor, CART and Random Forest models. Two sets of tuning parameters were explored for CART models, one which allowed for deeper trees, so five models are ultimately compared for performance. Obviously, linear regression assumes the model form of f(x) is linear which is helpful for inferential purposes but is sensitive to functional form mispecifications and is thus not usually

the best model for predictions. As is shown later on, the linear model performs far worse than its counterparts.

5. Results

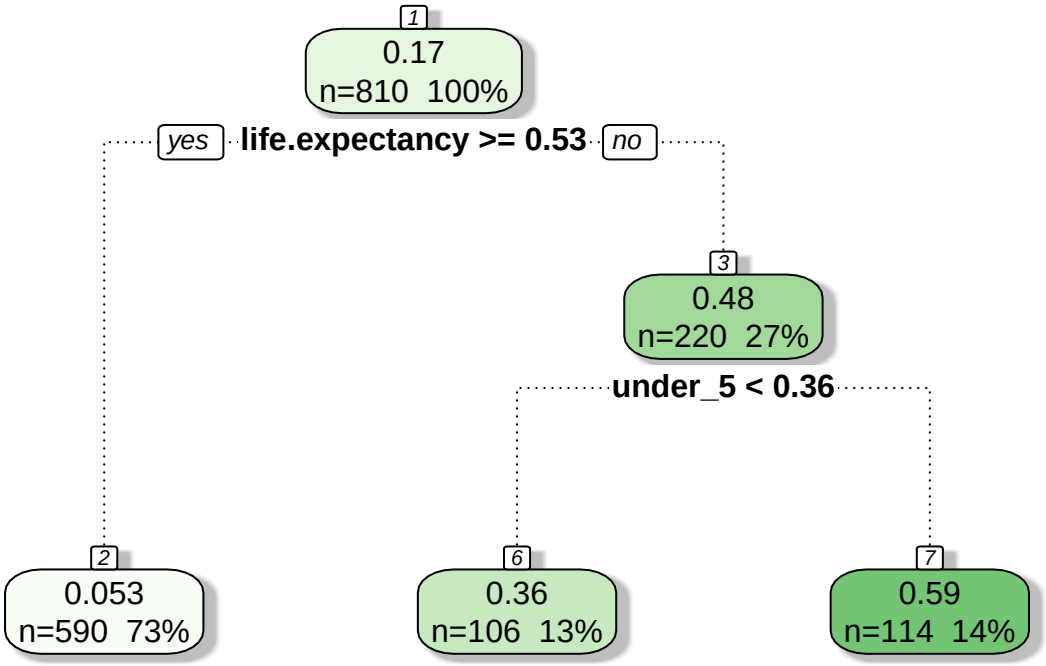
Unsuprsingly, the linear model performs the worst with am RMSE of .40 and an r-squared of .438. Priors and understanding of parametric analysis seemed to indicate that it would be the worst performing model. It is possible that the linear model performs very poorly on one of the folds causing large differences between the sample performance for each fold, creating larger variation and thus the larger confidence interval shown in the dotplot below. The classification and regression tree (CART) model yielded an RMSE of .119 and an R-squared of .734. K-Nearest Neighbor performed slightly worse with an RMSE of .122 and r-squared of .721 with an optimal model using k=7 out attempts k=5,7,9.

Expanding the grid and allowing trees to grow deeper yields only a slightly stronger CART model with an RMSE of .103 and an r-squared of .798. Finally, Random Forest is run, and using an mtry of 144, the final RMSE is .08 and R-squared is .866. Given that 135 countries are included in the dataset, which are turned from a categorical variable to a series of dummy variabes during pre-processing, a large mtry is logical. Random Forest was then run on the test data, producing an RMSE of .082, lower than the training data. The similarity in the two RMSE values indicates that the Random Forest model does well in estimating maternal mortality.

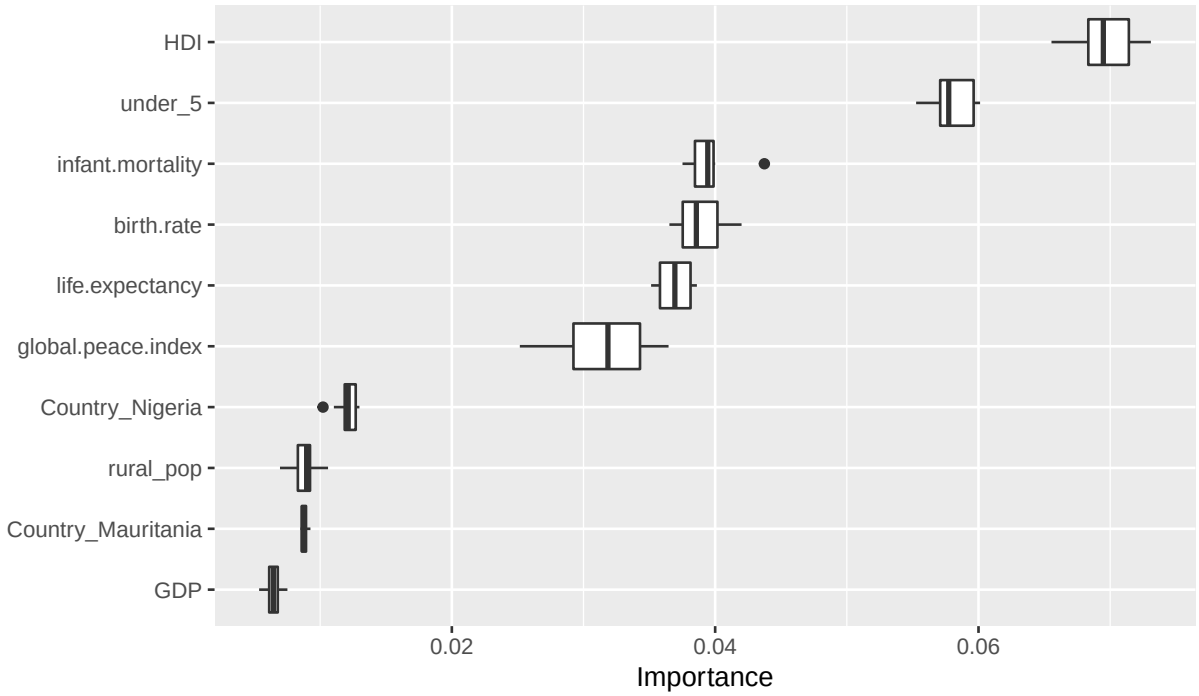


6. Discussion

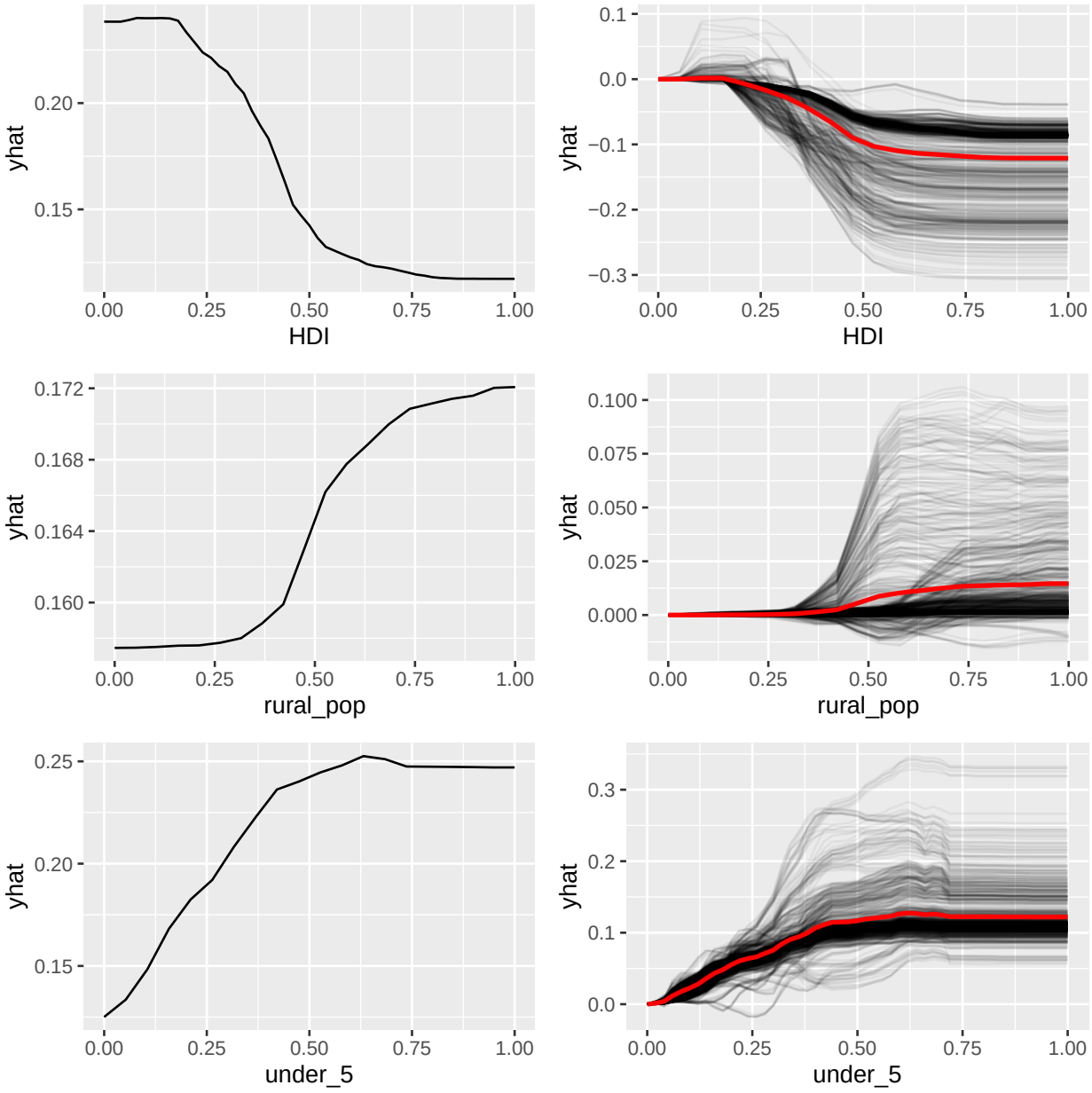
A visualization of the CART model helps highlight important factors when predicting mortality. Surprisingly, life expectancy equal to or above 0.53 on the normalized scale was the most significant predictor of maternal mortality. While the two may seem obviously correlated, available literature indicates that education and health access would be the most critical features to prediction. The Human Development Index is the second most important predictor, for countries with life expectancies below .53. Variable importance is discussed in more detail below.



The variable importance plot depicted below confirms much of the discussion in the literature previously discussed and hypotheses made about important features to predicting maternal mortality. Unsurprisingly, the Human Development Index variable is by far the most important in predicting maternal mortality. Perhaps more interestingly, the proportion of the population that is rurally located is among the least important feature to this model. Controlling for urban/rural location is a common feature in many public health models because it is strongly correlated with education, income, health access and many other socio-cultural factors that predict health outcomes. Additionally, health expense was one of the least important variables to the model. Health expenditure was included as a possibly proxy for health utilization. Given the empirical evidence that health utilization is an important predictor of maternal mortality, it is fair to conclude that this was not in fact a strong substitute measure. Interestingly, Nigeria and Mauritania were important variables to the Random Forrest model signaling that they had strong predictive power for Y, maternal mortality.



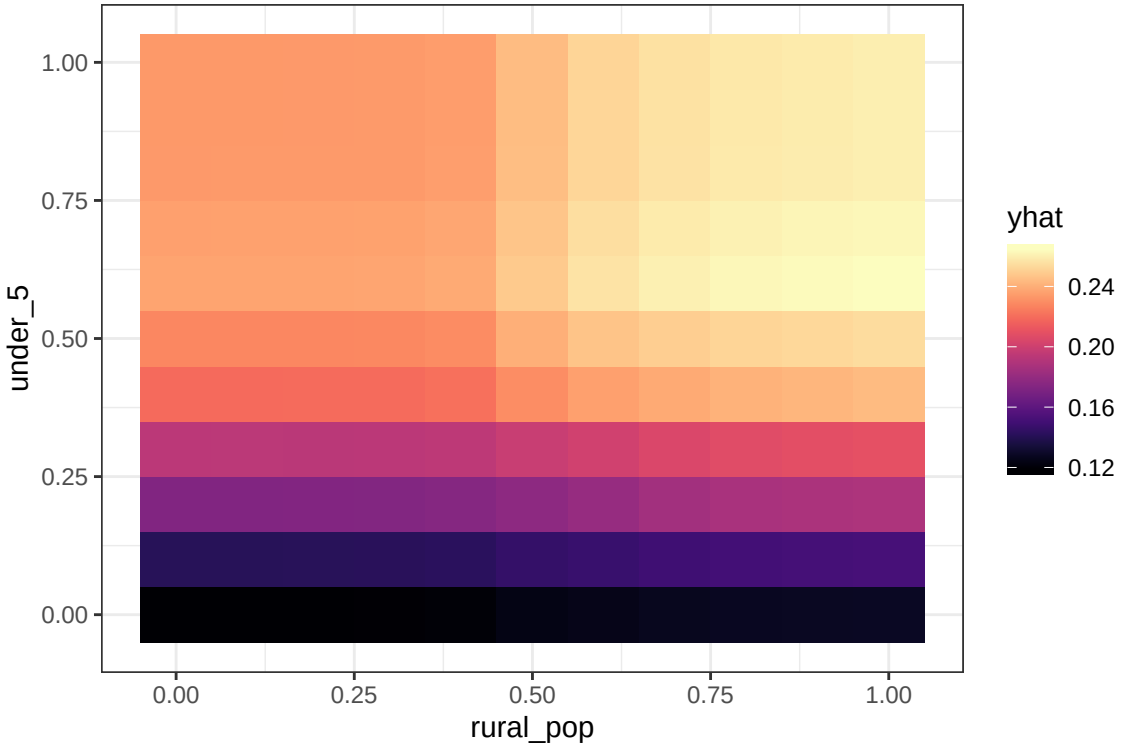
The partial effect of HDI, rural population and under-5 mortality on maternal mortality are shown below. At an HDI above ~.20, maternal mortality drops dramatically signaling some sort of threshold of development. Unsurprisingly, under-5 mortality has a steep, positive relationship with maternal mortality though, at ~.65 the slope levels suggesting that the effect of increasing under-5 mortality on maternal mortality diminishes above a certain point. The proportion of rural population is included, though not among the most significant variables because of the relationship it displays with maternal mortality. The proportion of rural population resembles an almost dichotomous relationship with maternal mortality above and below .50. Below .50 rural population, maternal mortality drops drastically and is relatively level among various proportions of rural populations. The same is true above .50.



Divergences in the individual conditional expectation (ICE) plots of rural population variable point towards possible interactions in the data. Similarly, the dramatic slope in the partial dependency plot for HDI indicated a threshold above ~.2 that signaled lower levels of maternal mortality but the ICE plots show that there's more heterogeneity in maternal mortality rates for countries with HDIs above that threshold that previously indicated.

A potential interaction between under-5 mortality and the proportion of the population that is rural is explored in an attempt to explain some of the heterogeneity ex-

hibited in the ICE plots. The thresholds of ~.50 for rural population still exists here but it is mitigated at the lowest levels of under-5 mortality. Meaning, even for countries with high proportions of the population living in rural areas, if under-5 mortality is low the maternal mortality rate is likely to be low as well. This is logical assuming many of the factors that drive under-5 mortality would predict maternal mortality as well.



7. Prediction

By design, publicly accessible data, particularly from highly visible sources such as the World Bank and World Health Organization, are likely to have been scoured and reanalyzed several times over. As a result, success for this project was not define by the novelty of the results. Success instead was evaluated based on the strength of the model, measured by r-squared and RMSE, as well as by how well the findings are explained to the audience. By these measures I hope the project is a success. The final mean squared error (MSE) of the Random Forest model using test data was .004.

I was surprised by the findings from my analysis of variable importance. Empirical evi-

dence suggests that residing in a rural location is highly significant in determining a number of socio-cultural outcomes including education attainment, health status, and income. It was the original hypothesis of this paper that a high proportion of the population living in rural areas would be a significant indicator of maternal mortality, for all the reasons outlined above. The variable importance plot reveals that it is not, at least relative to other variables. Further analysis can be undertaken to run this model at a district or country level to see if rural population becomes more significant when looking at intra-country differences. At the country-level it is possible that low maternal mortality in urban areas is masking maternal mortality in rural areas and decreasing the significance of that variable.

8. References

1. Beatriz, E. D., Molnar, B. E., Griffith, J. L., & Salhi, C. (2018). Urban-rural disparity and urban population growth: A multilevel analysis of under-5 mortality in 30 sub-Saharan African countries. Health & Place, 52, 196–204. doi: 10.1016/j.healthplace.2018.06.006

2. Birmeta, K., Dibaba, Y. & Woldeyohannes, D. Determinants of maternal health care utilization in Holeta town, central Ethiopia. BMC Health Serv Res 13, 256 (2013). https://doi.org/10.1186/1472-6963-13-256

3. GBD 2015 Maternal Mortality Collaborators (2016). Global, regional, and national levels of maternal mortality, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet (London, England), 388(10053), 1775–1812. https://doi.org/10.1016/S0140-6736(16)31470-2

4. Honaker, J. King, G. Blackwell, M. Amelia II: A Program for Missing Data. Journal of Statistical Software. (2011). 45(7), 1-47. URL http://www.jstatsoft.org/v45/i07/.

5. G. King and M. Blackwell, Amelia: Multiple Imputation. Harvard University, 2019.

6. Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1 https://CRAN.R-project.org/package=readxl

7. Hadley Wickham and Evan Miller (2019). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.2.0. https://CRAN.R-project.org/package=haven

8. Hadley Wickham and Lionel Henry (2019). tidyr: Tidy Messy Data. R package version 1.0.0. https://CRAN.R-project.org/package=tidyr

9. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr

10. Silge J, Robinson D (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS*, *1*(3). doi: 10.21105/joss.00037 (URL: https://doi.org/10.21105/joss.00037), <URL: http://dx.doi.org/10.21105/joss.00037>.

11. Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr

12. Hadley Wickham (2019). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.5. https://CRAN.R-project.org/package=rvest

13. Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

14. Brandon Greenwell, Brad Boehmke and Bernie Gray (2020). vip: Variable Importance Plots. R package version 0.2.2 https://CRAN.R-project.org/package=vip

15. Max Kuhn and Hadley Wickham (2020). recipes: Preprocessing Tools to Create Design Matrices. R package version 0.1.12. https://CRAN.R-project.org/package=recipes

16. Kimani-Murage, E., Fotso, J., Egondi, T., Abuya, B., Elungata, P., Ziraba, A., ... Madise, N. (2014). Trends in childhood mortality in Kenya: The urban advantage has seemingly been wiped out. Health & Place, 29, 95–103. doi: 10.1016/j.healthplace.2014.06.003

17. Ononokpono, D. N., & Odimegwu, C. O. (2014). Determinants of Maternal Health Care Utilization in Nigeria: a multilevel approach. The Pan African medical journal, 17 Suppl 1(Suppl 1), 2. https://doi.org/10.11694/pamj.supp.2014.17.1.3596

18. Schloerke, B. Crowley, J. Cook, D. Briatte, F. Et al. (2020).GGally: Extension to 'ggplot2'. R package version 1.5.0. https://CRAN.R-project.org/package=GGally