Christian Campbell **DSC 680 - Amirfarrokh Iranitalab** **11/14/2024**

Final Report

Topic:

Predicting and Analyzing Key Factors Influencing Underweight Women: A Machine Learning Approach with Feature Importance Analysis.

Business Problem:

I'm aiming to tackle a significant health issue: understanding and predicting what factors contribute to women being underweight, particularly women aged 15 to 49 with a BMI below 18.5. Undernutrition in women is not just a personal health concern—it impacts maternal and child health, leading to broader social and economic challenges. My goal is to uncover which environmental, geographical, and socio-economic factors are most influential in predicting this health status. By examining variables like urban versus rural location, altitude, rainfall, temperature, economic conditions (such as poverty and food prices), and accessibility to urban centers, I hope to reveal patterns and insights that can guide public health strategies. This analysis can help policymakers and aid organizations target interventions more effectively, ultimately supporting efforts to reduce malnutrition and improve the health outcomes of women and their families.

Background/History:

Malnutrition is a significant concern among women in Sub-Saharan Africa, particularly for those aged 15 to 49. This issue is not only a personal health challenge but also affects maternal and child health, with wider social and economic consequences. Despite their physically demanding lives and limited economic resources, the prevalence of malnutrition among women, though severe, is not as extensive as one might expect given these challenges.

Various studies, including the United Nations' assessments, have highlighted that around 20% of women in this region have a BMI below 18.5, indicating chronic energy deficiency. This figure is comparable to other developing regions but lower than in areas like South Asia, where nearly 60% of women are underweight. Other indicators, such as height and arm circumference, reinforce that while malnutrition is present, the region's nutritional status varies.

Indicators such as maternal mortality and low birth weight percentages also reveal the extent of malnutrition and its implications. For example, Western and Middle Africa show higher maternal mortality rates, often linked to nutritional deficiencies. While some countries in Eastern and Southern Africa fare better, food insecurity remains a significant issue throughout the region.

The nutritional status of women is influenced by factors beyond food availability, including socioeconomic conditions, access to healthcare, and environmental factors. By understanding these influences, targeted public health interventions can be more effectively designed to combat malnutrition and support women's health and well-being.

Christian Campbell **DSC 680 - Amirfarrokh Iranitalab** **11/14/2024**

Final Report

Data Explanation:

I'm using a dataset from Kaggle that focuses on 11 priority countries identified by the USAID Feed the Future (FTF) initiative: Bangladesh, Ethiopia, Ghana, Guatemala, Honduras, Kenya, Mali, Nepal, Nigeria, Senegal, and Uganda. The dataset includes a variety of features, such as whether the area is urban or rural, altitude, rainfall estimates, country-specific indicators, climate variables like land surface temperature, and geographic details like latitude and longitude. It also encompasses data on socio-economic and environmental factors such as poverty levels, significant events like conflicts, and market food prices, which are essential for understanding local economic conditions and food security. The data selection is designed to include publicly available information, making it practical for use by organizations that might not have the resources for extensive data collection. This approach allows for a comprehensive analysis that considers location, climate, vegetation, and socio-economic elements, all of which are known from previous studies to influence patterns of poverty and malnutrition.

I started by removing the first column, which was just an ID, and dropped the market columns since they had too many missing values to be helpful. For columns with a small percentage of missing data (like "chrps" at 1.5%, "lst" at 4%, "sif" at 1.6%, and "underweight_bmi" at 1.4%), I filled in the missing cells using the average of each column. I also changed the "country" column into numerical values for easier processing, assigning numbers to each country.

Methods:

For my project, I used a machine learning approach to analyze the data and predict the key factors that influenced whether women were underweight. I trained a random forest regressor model, which was effective at handling complex data and capturing important relationships. To optimize the model's performance, I used a technique called GridSearchCV to fine-tune its settings. Once the model was trained, I examined which features were most important for making predictions by using the feature importance tool from the random forest model. To gain deeper insights, I applied SHAP (Shapley Additive Explanations) values, which explained how each feature impacted individual predictions in detail. I created visualizations such as bar charts and SHAP summary plots to clearly present this information. To evaluate the performance of my model, I checked various metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2).

Analysis:

To evaluate my random forest regressor model, I tested many different parameter settings (972 combinations in total) to find the best one. The final model had parameters like using 200 decision trees and allowing full tree depth with specific rules for splitting and minimum leaf size. The model performed well on the training data with a Mean Absolute Error (MAE) of 0.0361 and a high R-squared value of 0.8742, indicating it explained about 87% of the data's variance. On the test data, the MAE was 0.0732, and the R-squared dropped to 0.5295, suggesting the model could explain about 53% of the variability for new data. The most important features affecting predictions were longitude, country, latitude, travel time to urban centers, and the prevalence of wasting in women. Other factors like altitude, economic indicators, and various climate and geographic features also played significant roles.

Conclusion

In conclusion, my project aimed to shed light on the complex issue of underweight women in certain developing countries by using machine learning to identify key contributing factors. The random forest model I trained helped uncover which aspects—such as geography, economic conditions, and environmental factors—played the most significant roles in predicting whether a woman would be underweight. The results highlighted that features like longitude, country, and latitude had the most influence, while other factors such as travel time to urban areas, altitude, and socio-economic indicators like poverty also showed considerable impact. By identifying these influential variables, this analysis could provide valuable insights for public health officials and aid organizations looking to target their resources more effectively. Ultimately, the goal is to inform public health strategies that improve women's nutrition and overall well-being, benefiting families and communities at large.

Assumptions

Several assumptions were made during the project to simplify and focus the analysis. First, I assumed that the dataset was comprehensive and representative of the population within the countries studied. I also assumed that filling in missing values with averages would not significantly distort the data's true distribution. The numerical conversion of categorical variables, such as countries, assumed that this approach would allow the model to learn patterns effectively. Finally, I assumed that external factors like government policies and food programs remained constant and did not significantly skew the results.

Limitations

This project had its share of limitations. One significant limitation was the quality and completeness of the data. While I handled missing values by filling them in with averages, this approach could oversimplify the data and potentially hide nuanced trends. Additionally, the dataset was limited to 11 specific countries, which means the model’s applicability might be restricted to those regions or ones with similar conditions. The complexity of real-world factors influencing nutrition—such as sudden economic shifts, conflicts, or public health crises—was not accounted for in the model. Lastly, machine learning models, even well-tuned ones like my random forest, are not immune to biases in the data, which could affect the generalizability of the results.

Challenges

Throughout the project, I faced several challenges. The first major hurdle was data preparation, particularly dealing with missing or incomplete records. I had to make careful decisions about what to keep, drop, or fill in, which took time and could influence the final results. Another challenge was fine-tuning the model to find the best parameters. This required running GridSearchCV over many combinations, which was time-consuming. Interpreting the importance of features and translating those insights into actionable information was also challenging, as the relationships between variables like climate, location, and socio-economic status were often intertwined.

Future Uses/Additional Applications

The findings and methodology from this project could be expanded for broader real-world applications. Public health agencies and NGOs can use these insights to target nutritional aid and resources more effectively, directing them to areas most at risk based on the factors found important in my analysis. Governments and policymakers could implement this approach to create more data-driven interventions, tailoring programs to specific regions' unique challenges. Additionally, the framework could be adapted to analyze other public health issues, such as predicting child stunting or the effects of economic policies on nutrition. The adaptability of machine learning models like the random forest makes it possible to use them in related applications involving socio-economic and environmental data.

Recommendations

For future projects or extensions, I recommend incorporating additional variables that could offer deeper insights, such as healthcare access, education levels, and local economic policies. Ensuring data quality should be a priority; collecting more comprehensive and up-to-date datasets would enhance the model's accuracy and reliability. Collaborating with public health experts could also provide context to interpret the findings more effectively and align them with actionable strategies. I also suggest exploring other machine learning algorithms to compare their performance and see if any improvements could be made over the random forest regressor.

Christian Campbell **DSC 680 - Amirfarrokh Iranitalab** **11/14/2024**

Final Report

Implementation Plan

To implement a similar project in a real-world setting, first I would establish partnerships with local government agencies, NGOs, and health organizations to access relevant and comprehensive datasets. Next, a robust data collection and cleaning phase would be critical to ensure the dataset's accuracy and representativeness. Once prepared, the data would be used to train and validate machine learning models, with domain experts providing input to refine feature selection and model tuning. After running the analysis, findings would be presented through reports and visualizations, highlighting actionable insights. Finally, public health officials and community planners could use these insights to guide interventions, monitor the impact, and adjust strategies as needed. This collaborative, data-driven approach could significantly contribute to tackling malnutrition among women and improving overall public health.

Ethical Considerations:

When working on a project like this, it's important to consider how sensitive the topic is and what potential impacts it could have on the people involved. One major ethical concern is ensuring that the data is used responsibly and in a way that doesn't stigmatize or negatively impact the communities it represents. For example, labeling certain areas or groups as more prone to undernutrition could reinforce negative stereotypes or lead to policies that aren't well thought out. There's also the risk of using data that might not fully capture the personal and cultural factors affecting individuals' health, which could oversimplify complex issues. Additionally, data privacy is crucial, so even though this dataset is public, I need to make sure that any insights drawn don't unintentionally expose or misinterpret personal or community information. The goal is to provide helpful, unbiased insights that can support positive change without causing harm or unfair treatment.

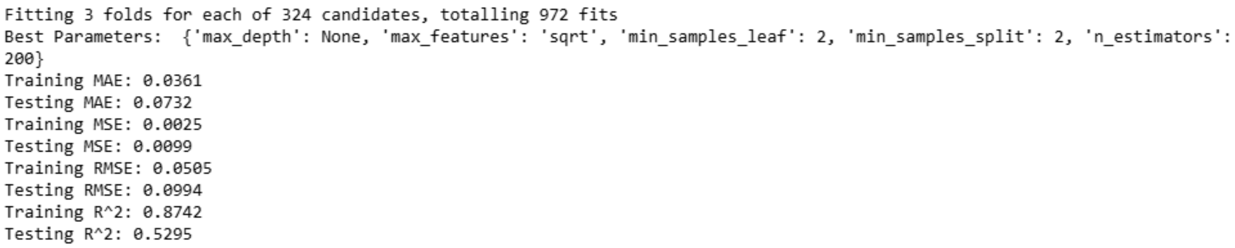
References:

- Browne, C., Matteson, D. S., McBride, L., Hu, L., Liu, Y., Sun, Y., Wen, J., & Barrett, C. B. (2024, January 2). *ML prediction of poverty and malnutrition dataset*. Kaggle. https://www.kaggle.com/datasets/adamprzychodni/ml-prediction-of-poverty-and-malnutrition-dataset?select=Appendix.pdf
- Browne, C., Matteson, D. S., McBride, L., Hu, L., Liu, Y., Sun, Y., Wen, J., & Barrett, C. B. (2021). Multivariate random forest prediction of poverty and malnutrition prevalence. *PLOS ONE*, *16*(9). https://doi.org/10.1371/journal.pone.0255519
- U.S. National Library of Medicine. (1996, January 1). *Nutritional status*. In Her Lifetime: Female Morbidity and Mortality in Sub-Saharan Africa. https://www.ncbi.nlm.nih.gov/books/NBK232554/

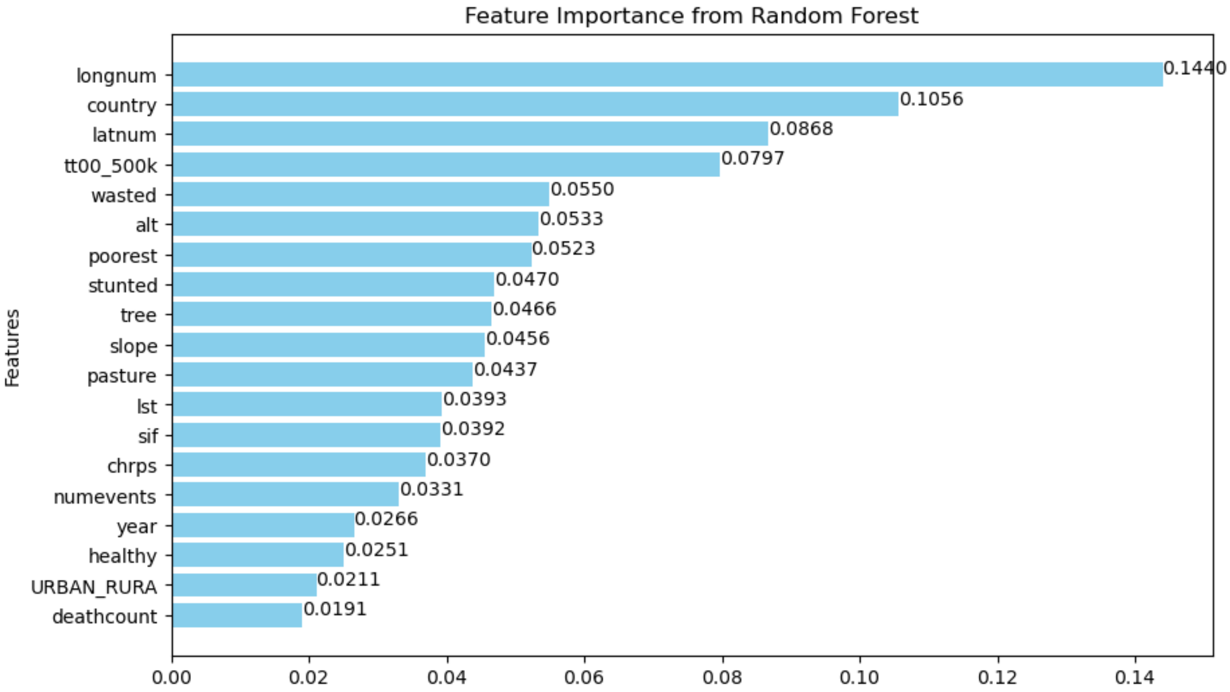
Final Report

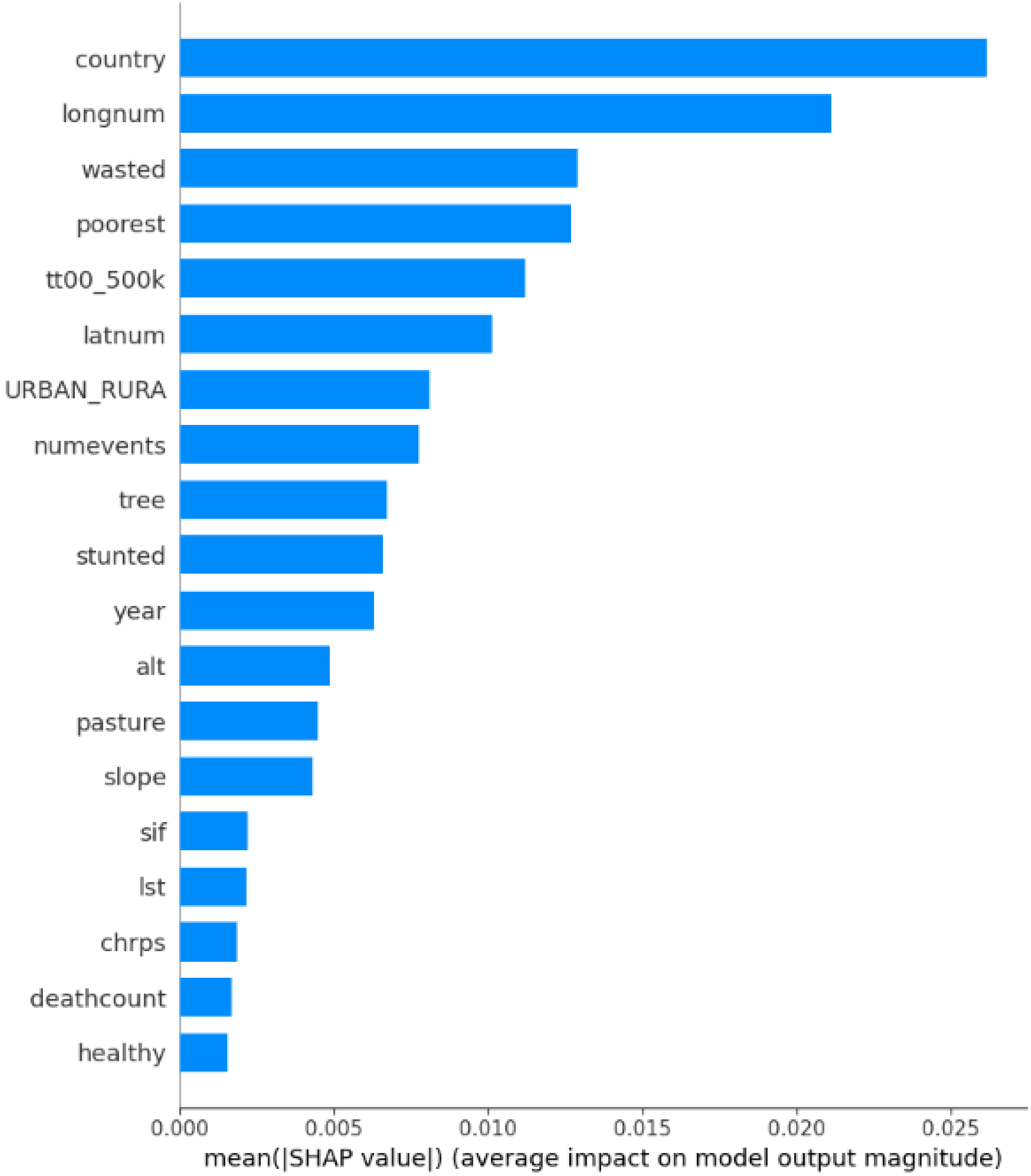
Illustrations:

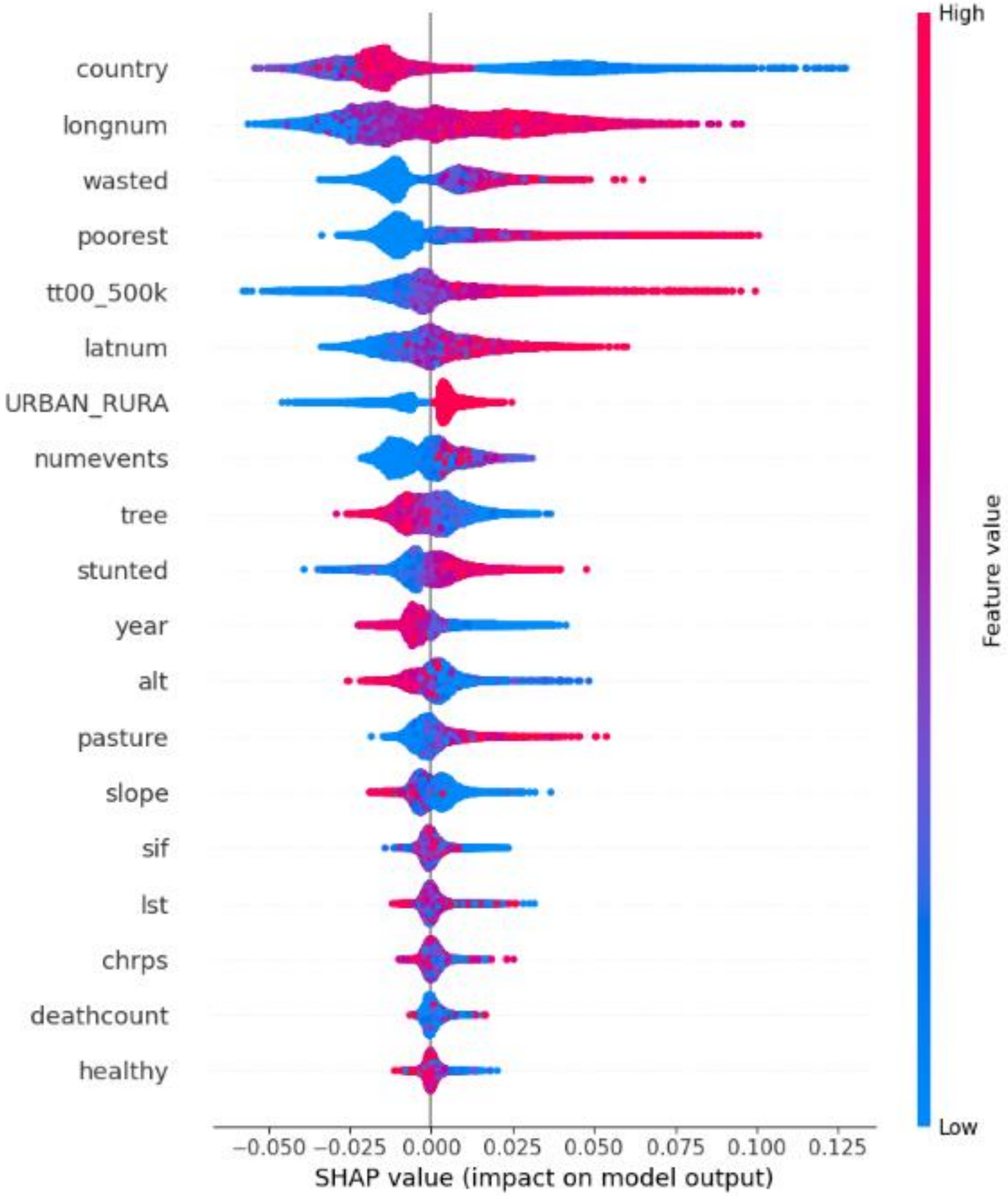
Evaluation Results



```
Fitting 3 folds for each of 324 candidates, totalling 972 fits
Best Parameters:  {'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators':
200}
Training MAE: 0.0361
Testing MAE: 0.0732
Training MSE: 0.0025
Testing MSE: 0.0099
Training RMSE: 0.0505
Testing RMSE: 0.0994
Training R^2: 0.8742
Testing R^2: 0.5295
```




country
longnum
wasted
poorest
tt00_500k
latnum
URBAN_RURA
numevents
tree
stunted
year
alt
pasture
slope
sif
lst
chrps
deathcount
healthy
0.000
0.005
0.010
0.015
0.020
0.025
mean(|SHAP value|) (average impact on model output magnitude)


country
longnum
wasted
poorest
tt00_500k
latnum
URBAN_RURA
numevents
tree
stunted
year
alt
pasture
slope
sif
lst
chrps
deathcount
healthy
−0.050
−0.025
0.000
0.025
0.050
0.075
0.100
0.125
SHAP value (impact on model output)
High
Feature value
Low