Assignment 1: Introduction to ML

Sam Peterson s3722250

9th April 2024

0.1 Data Transformation and Model Training Overview

- The process of transforming the data and training the final model involved several key steps:

0.2 Baseline Model:

- The baseline model selected was a linear regression model. It was chosen as during the EDA multiple variables were visualized with having linear relationships in the scatterplot graphs along with high correlation values with the target variable. As the dataset includes multiple features, multivariate regression was chosen over univariate. No preprocessing steps had been undertaken on the data before the baseline model had been trained.

Multivariate linear regression hypothesis.

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

- The hypothesis equation calculates the predicted value theta(x) by taking a weighted sum of the features x_1, x_2, ..., x_n with their respective coefficients theta_1, theta_2, ..., theta_n, and adding the bias term theta_0.

0.3 Evaluation Metrics

1. **Mean Squared Error (MSE):**
   - MSE provides a measure of how close the predicted values are to the actual values. Lower MSE indicates better model performance in terms of prediction accuracy.
   - Squaring the errors in MSE emphasizes larger errors, making it more sensitive to outliers and deviations (Which EDA has shown is prevalent in the data).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

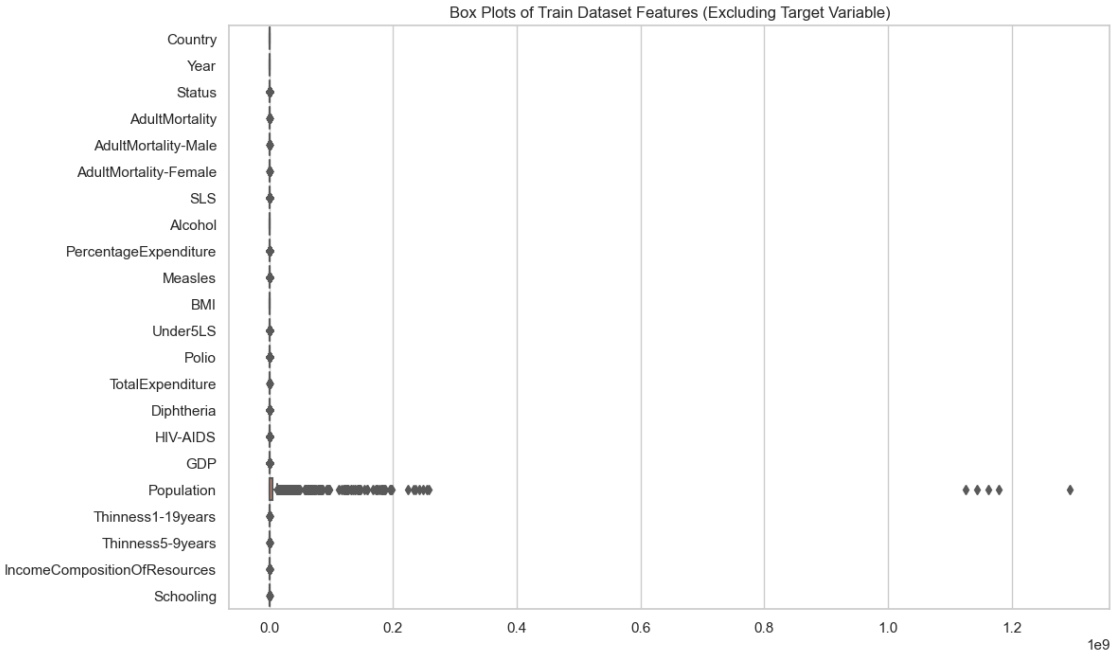
2. **R-squared (Coefficient of Determination):**
   - R-squared measures the goodness of fit of the regression model. A higher R-squared value indicates that more variance in the target variable is explained by the features.
   - It helps assess the proportion of variability in the target variable that can be attributed to the model's predictions, providing insights into model effectiveness.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

0.4 Data Splitting and Validation Techniques.

- During the training of the initial models and during the incremental data preprocessing steps, the Hold Out method of data splitting was implemented. With an 80% training and 20% testing split in the data. This ratio was chosen as it provides a good balance between the size of the training set and the validation set. The Holdout Method is straightforward to implement and computationally efficient. A limitation of the Holdout Method is that the performance estimate may be sensitive to the specific random split of the data. All the preprocessing techniques were performed on the training and validation sets separately to ensure that no data leakage would occur. This is true for all splitting techniques implemented. Later in the development of the model, when the final two models were settled upon. A combination of cross validation and hyperparameter tuning took place. This was to find the optimal performing model on average over 5 training/validation splits for each value of alpha. Whilst this took place every training and validation set generated underwent preprocessing separately in order to ensure no data leakage would happen throughout this process.

Figure 1. Boxplots of all the independent variables.



- From *Figure 1* we can see that the scale difference between the variable values is apparent. Especially between "Status' ' and"Population". The difference between means of these two variables is 6.492×10ˆ7 in magnitude. The significant scale difference highlights the importance of scaling in preprocessing. Scaling features ensures that all variables contribute equally to the model's predictions, preventing features with larger magnitudes from dominating the

training process. Therefore, scaling is crucial for achieving a balanced and effective regression model.

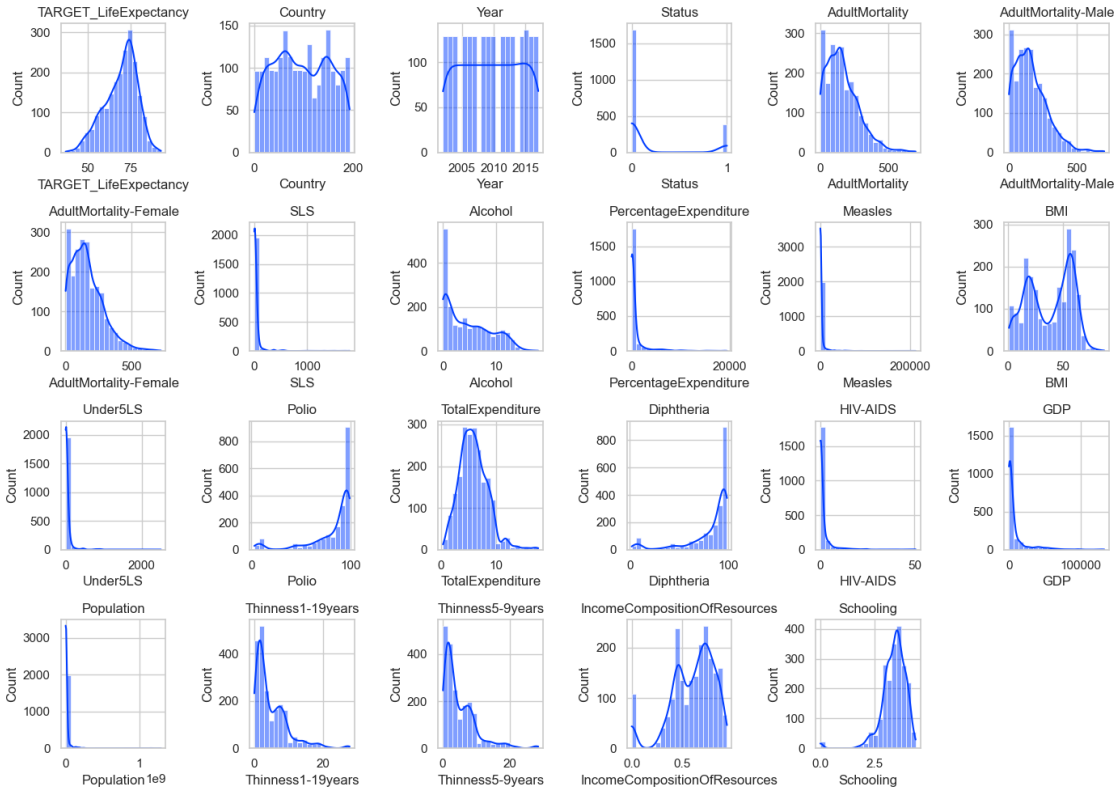
0.5 Outlier Handling

- It was decided not to perform outlier removal or transformation. If we were to remove all the outliers based on Tukey's rule, we would lose 1427 values. The loss or alteration of that many data points would only serve to negatively impact the performance of the model so it was not performed during the pre-processing. Understandably, the performance of the model might have suffered as a result of this choice. However, without access to more sophisticated outlier handling techniques it was chosen not to risk the removal of potentially information rich data points at the expense of potential noise.

0.6 MinMax Scaling:

- The initial step in preprocessing involves applying MinMax scaling to address the large variation in scale observed among the independent variables during exploratory data analysis (EDA). MinMax scaling transforms the features to a common scale, typically between 0 and 1, thereby preventing features with larger ranges from dominating the model and aiding in comparing the importance of different features based on their magnitudes.

Figure 2. Histogram plots of all variables.



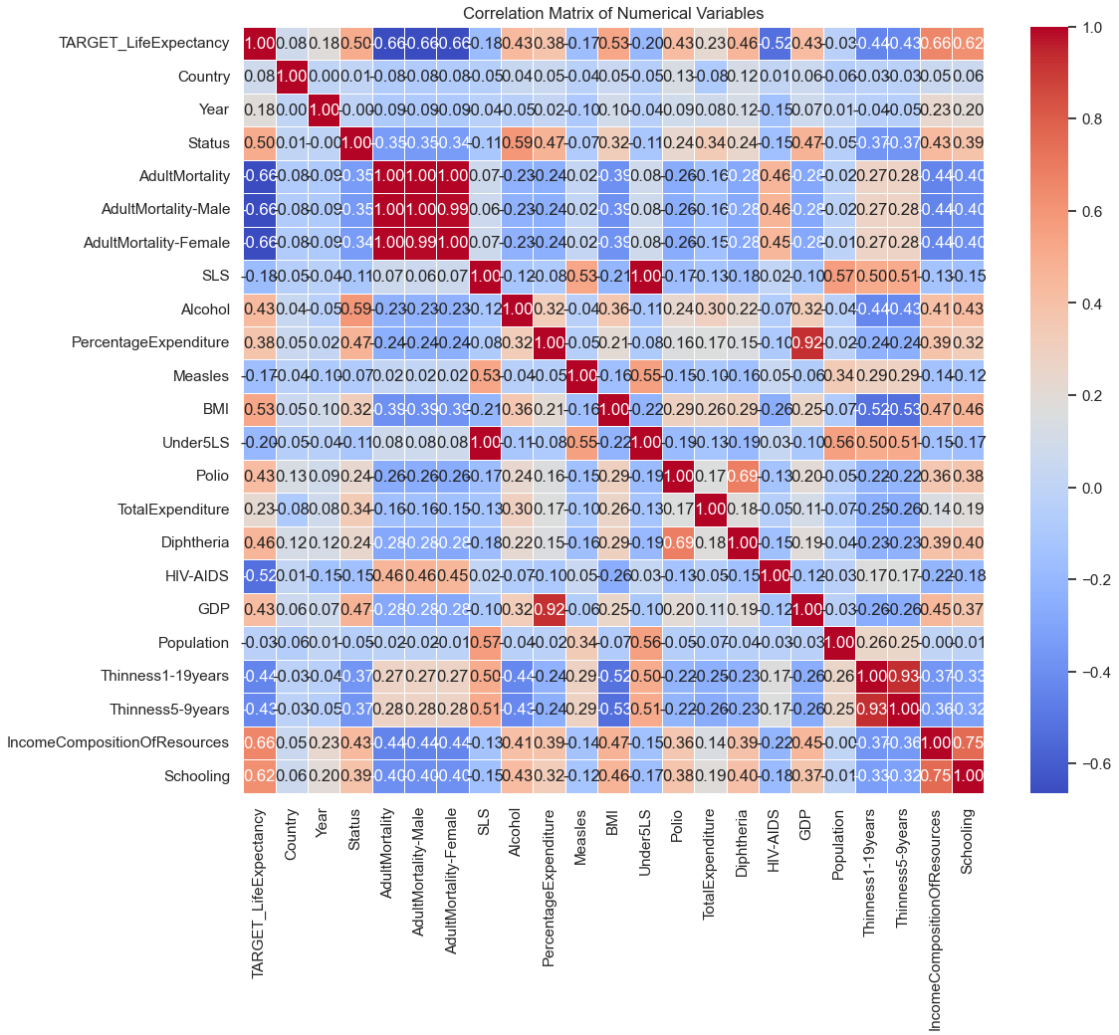
0.7 Yeo-Johnson Normalization:

- To handle the skewed distributions observed in the histograms during EDA, Yeo-Johnson normalization was applied. This transformation was chosen for its robustness to outliers and its ability to handle zero and negative values. By applying techniques like Yeo-Johnson normalization, we ensure that the data is more evenly distributed, making it easier for the model to learn meaningful patterns and relationships. This ultimately leads to more accurate predictions and better generalization to unseen data.

0.8 Categorical Variables Handling:

- Given the high variation in categories observed in the 'Country' and 'Year' variables, methods such as one-hot encoding were deemed suboptimal due to the curse of dimensionality.

Figure 3. Correlation matrix and heatmap of all variables within the training dataset.



0.9 Regularization Techniques:

- To further improve the model's performance and address potential overfitting, Ridge and Lasso regularization techniques were introduced. These techniques help by controlling the model's complexity and reducing the impact of multicollinearity, which can lead to more robust and generalizable predictions. Ridge regularization (L2) encourages smaller coefficients, while Lasso regularization (L1) performs feature selection by shrinking less important features' coefficients to zero. The aim of incorporating these regularization techniques is to enhance the model's ability to capture the underlying patterns in the data while avoiding overfitting and improving its interpretability. Due to the high dimensionality of this dataset Lasso regularization may serve us as it can reduce the impact of the features with low correlation to the target variable.
  - L1 Regularization (Lasso) L1 regularization, also known as Lasso (Least Absolute Shrinkage and Selection Operator), adds a penalty term to the loss function that is proportional to the absolute value of the coefficients:

$$\text{Loss} = \text{Original Loss} + \lambda \sum_{i=1}^{n} |\beta_i|$$

  - L2 Regularization (Ridge) L2 regularization, also known as Ridge regression, adds a penalty term to the loss function that is proportional to the squared magnitude of the coefficients:

$$\text{Loss} = \text{Original Loss} + \lambda \sum_{i=1}^{n} \beta_i^2$$

0.10 Polynomial Model Adoption

- As the Regularisation steps were not seeing much of a change in the evaluation metrics performance, it may be an indicator that our linear hypothesis is not able to catch non-linear relationships present in the data. In order to test this, the preprocessed data was subsequently trained on a polynomial regression model.

In polynomial regression, the hypothesis equation for predicting the target variable (y) based on a polynomial function of the input feature (x) can be written as:

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + ... + \theta_n x^n$$

0.11 Hyperparameter Tuning and Cross Validation:

- Hyperparameter tuning, specifically the alpha value for Lasso and Ridge regularization, was performed using GridSearchCV to find the optimal balance between overfitting and underfitting. Additionally, 5-fold cross-validation was employed to ensure robustness and generalization of the model's performance across different data subsets.

Table 1. Iterative development of the final Model displaying the mse and R^2 error for each step.

Evaluation Metrics	Mean Squared Error	R-squared Score
Baseline	23.7599	0.7124
After MinMaxScaling	23.7599	0.7124
After Yeo-Johnson Normalization	18.2294	0.7794
Ridge Regularization	18.4323	0.7769
Lasso Regularization	18.6141	0.7747
Polynomial Regression	10.9843	0.8670
Ridge Regularization on Polynomial	10.4736	0.8732
Lasso Regularization on Polynomial	11.8473	0.8566
Mean tuned CV-5 Ridge-Poly Model Alpha(300)	15.8682	0.8209
Mean tuned CV-5 Lasso-Poly Model Alpha(0.1)	14.4743	0.8365
(Parameters chosen for final model)		
FINAL MODEL Lasso Regularization on Polynomial Regression Alpha(0.1)	10.7058	0.8704

0.12 Final Model Selection

- The Lasso Regularized Polynomial Regression Model, with an optimal alpha value of 0.1, emerges as the most effective among the developed models. With an average R-squared score of 0.8365 and a mean squared error (MSE) of 14.4743, it demonstrates a strong ability to predict the target variable based on the features. Leveraging polynomial regression, the model captures non-linear relationships within the data. Through regularization, cross-validation, and hyperparameter tuning, it underwent rigorous optimization. The final model, denoted as Lasso(alpha=0.1), is well-equipped to make reliable predictions on unseen data.

0.12.1 Limitations and Real-World Applications:

- While the developed model demonstrates promising performance, several limitations must be acknowledged for real-world applications:
- Data Abundance: In real-world scenarios, having a larger dataset would be preferable to increase confidence in the model's predictive power and its ability to generalize to unseen data.
- Outlier Handling: More sophisticated outlier handling techniques, such as z-score removal, is necessary to deal with outliers effectively.
- Feature Selection: Given the high dimensionality and potential redundancy in some variables, feature selection techniques would need to be implemented to improve model interpretability and reduce overfitting.
- Categorical Variable Handling: Better methods for handling categorical variables, such as binning techniques to reduce the number of categories or obtaining additional information about the country codes, would be beneficial to improve model performance.

In summary, while the developed model shows promise, further refinement and consideration of these limitations would be necessary for successful deployment and application in real-world scenarios.