Solution Report
On

Prediction of Life Expectancy

By
Keegan Fernandes
Karan Punjabi
Reena
Anujit Ghosh
Team 1

HackerVerse 2022

Department Of Data Science
Christ (Deemed to be University), Lavasa

Solution Document

Problem statement

Develop a life expectancy predictive model.

Data cleaning and Preprocessing

We identified many discrepancies in the data with missing values and poor column formatting.

Firstly, we identified the number of null values for each attribute and followed a different approach for columns with missing values greater than 50. Except Life_Expectancy, missing values from columns with less than 50 missing values were replaced with the interpolated value of the respective column. Data where Life_Expectancy was null, was removed. The other missing values were interpolated according to the respective country. The remaining missing value rows were removed from the dataset.

A few rows contained nonsensical data such as adult mortality rate being more than thousand in India when the data is out of 1000. We removed all such rows in the dataset

Once the data was void of null values, the next step was to remove outliers. The number of outliers and its percentage in each variable was calculated using Turke's method. The data set was normalized by winsorizing the necessary columns.

Once all necessary columns were normalized, the new winsorized columns were added to the dataset. All further analysis was done on the normalized data.

While normalizing the dataset, we found that Measles was affected by the normalization by 19%, raising a red flag that this may affect the models regressively. Therefore, the ‘Measles’ attribute was dropped.

Observations for any correlations among independent attributes helped identify and remove unnecessary attributes. A heat map was used to support the decision as evidence of correlation among independent attributes.

winsorized_Schooling, winsorized_Income_Comp_Of_Resources, winsorized_HIV, and winsorized_Adult_Mortality are evidently correlated with winsorized_Life_Expectancy.

However, it is also to be noticed that winsorized_Income_Comp_Of_Resources is correlated with winsorized_Schooling. As features should NOT be interrelated, one of the attributes must be eliminated. The attribute with a lower correlation to winsorized_Life_Expectancy

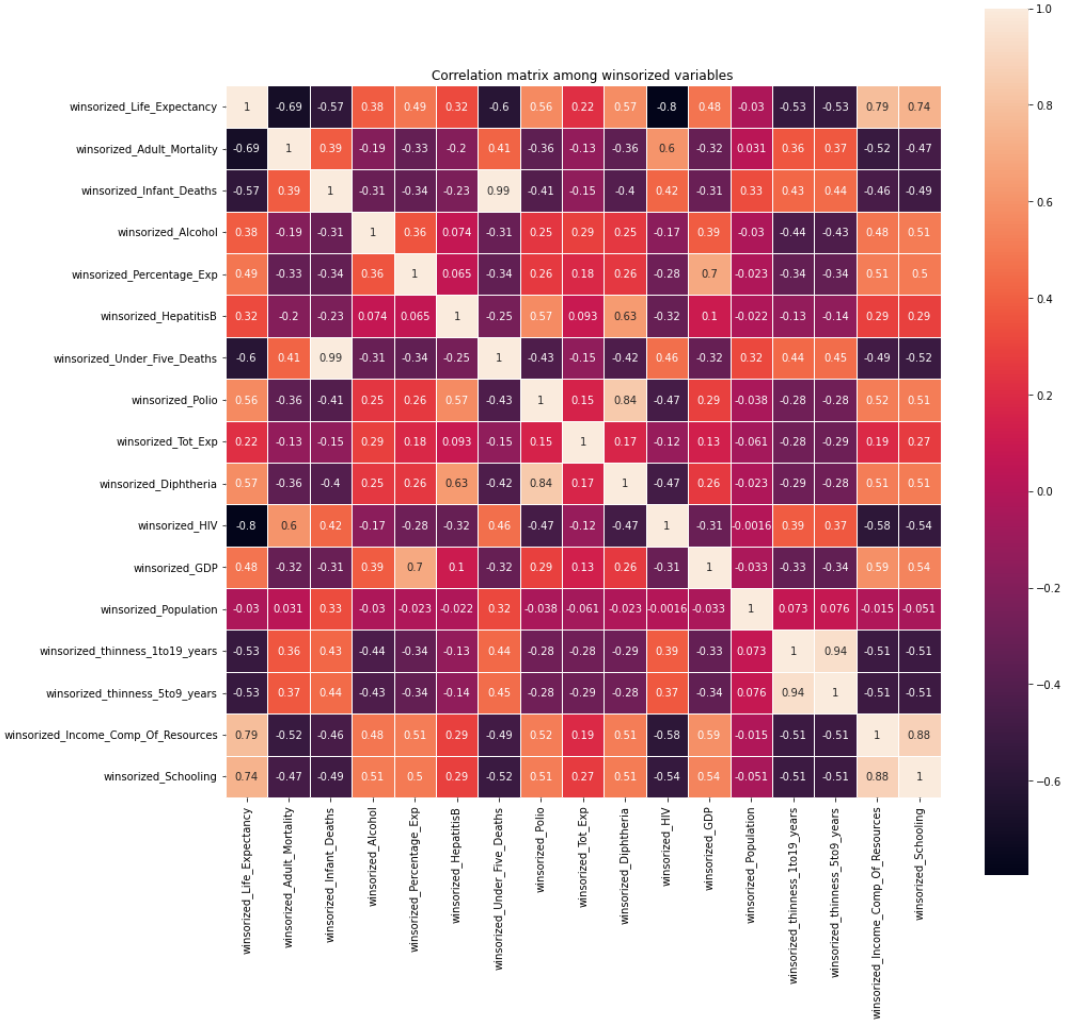
(winsorized_Schooling corr: 0.76; winsorized_Income_Comp_Of_Resources:0.82) was eliminated.

The attribute 'HIV' is bound to cause a bias in the prediction because the countries where HIV is prevalent i.e., African Countries are also the countries with low life expectancy. This factor does not necessarily directly contribute to life expectancy.

Many machine learning models assume the target variable to be normally distributed. If normality is an assumption non-normal variables can be transformed into normally distributed features using transformations such as taking the square root, the inverse, or the log of the variable, or by applying the Box-Cox transformation.

EDA

We explored the relation between variables to help decide the attributes for training the model.



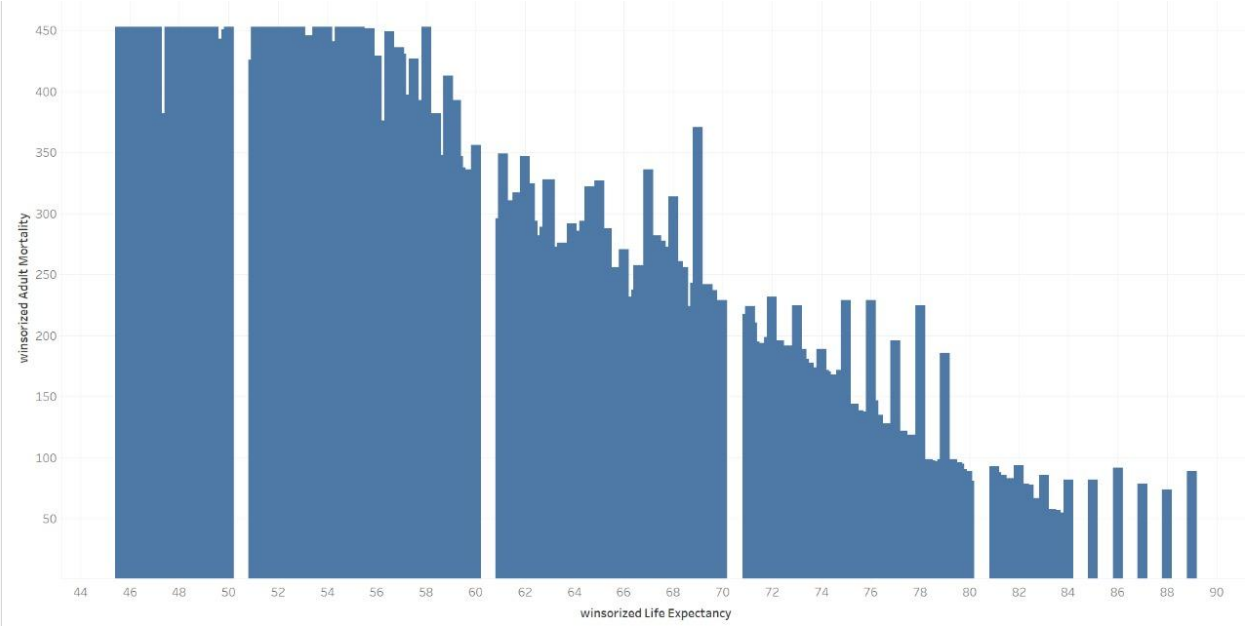
First, we attempted to understand the correlation between attributes for further investigation. Observations from the correlation matrix are given below**:**

- winsorized_Income_Comp_Of_Resources is correlated with winsorized_Schooling.
- winsorized_thinness_1to19_years and winsorized_thinness_5to9_years are correlated.
- winsorized_Polio and winsorized_Diphtheria are correlated.
- winsorized_Percentage_Exp and winsorized_GDP are correlated.
- winsorized_Income_Comp_Of_Resources and **winsorized_Life_Expectancy** are correlated.
- **winsorized_Life_Expectancy** and winsorized_Schooling are correlated.
- **winsorized_Infant_Deaths** and **winsorized_Under_Five_Deaths** are correlated.
- winsorized_HIV has a negative relation with **winsorized_Life_Expectancy**.
- winsorized_Adult_Mortality has negative relation with **winsorized_Life_Expectancy**

From the Heat map, we see that there are some features that are not strongly correlated to winsorized_Life_Expectancy. Therefore, we can say that all the features available in the dataset do not really affect the Life expectancy. winsorized _Adult_mortality,winsorized _Schooling, winsorized_Diphtheria, winsorized_Percentage_Exp, winsorized_Infant_Deaths, winsorized_GDP, winsorized_Income_Comp_of_Resources, winsorized_Tot_Exp, winsorized_thinness_5to9_Years, winsorized_thinness_1to19_Years are the columns that have a strong correlation with Life_Expectany.

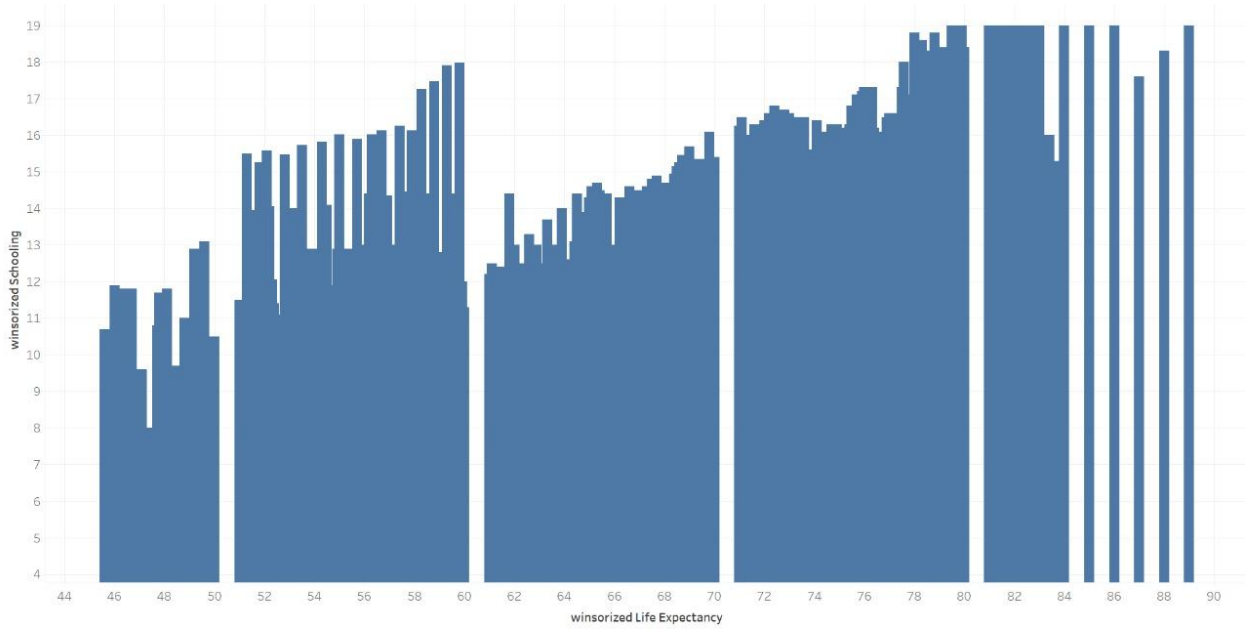
From a z test applied on countries having a life expectancy value less than 65 against countries having life expectancy greater than the average, we found that increasing the Percentage_Exp does not affect life expectancy. However, according to the dataset, this might also be due to faulty data. For example Afghanistan and Algeria have high life expectancy but this might be due to bad data collection.

The graph below shows the relation between Life_Expenctancy and Adult_Mortality. Life expectancy is inversely proportional to adult mortality i.e., negative correlation.

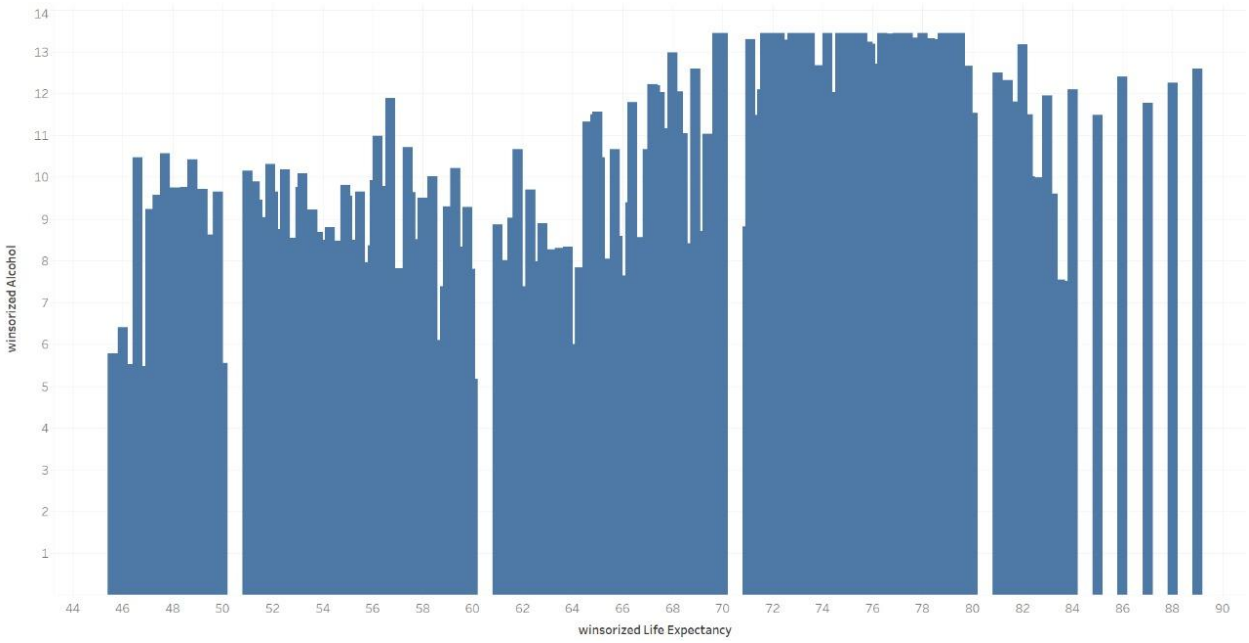


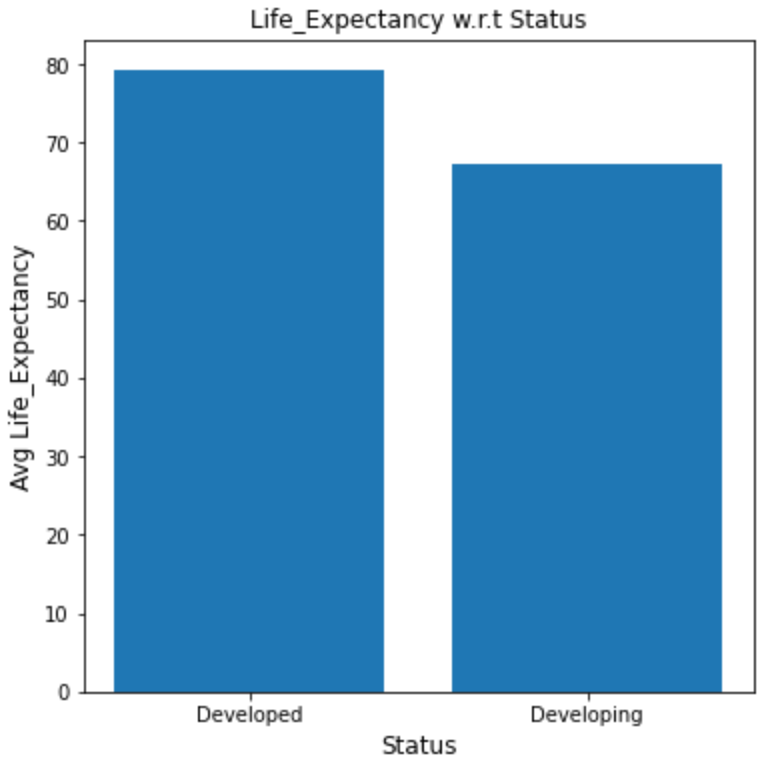
The graph below shows the relation between Life_Expenctancy and Infant_Deaths. Life expectancy is inversely proportional to infant deaths i.e., negative correlation.

The graph below shows the relation between Life_Expenctancy and Schooling. Life expectancy gradually increases as schooling increases. There is a slight positive correlation.



There is no relation between life expectancy and alcohol consumption. To further back this claim, we can look at the heat map, between windsorized_Life_Expectancy and Windsorize_Alcohol. The correlation is 0.38 which further points towards the same fact.





Feature Extraction

We used Feature importance to extract the most important features from the dataset. We did not include features like HIV which were giving us overfitted results.

Approach

a. **Random Forest Regression:** This model compiles many model's predictions to create a more accurate model. We decided to create a baseline model with Random Forest Regression because there were not many models that explored this approach.

b. **Ridge Regression:**

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated. It has been used in many fields including econometrics, chemistry, and engineering.

c. **Xgboost:**

XGBoost is a powerful model and is a standard model for non-neural networks.

Test

1. We first used RandomForestRegressor on all the features and got an accuracy of 0.895651449 and an R^2 score of 0.849044387. We then used Ridge Regression on all the features and got an accuracy of 0.0.999999983 and an R^2 score of 0.0.999999983.
2. Due to the high score, we obtained we realized that some features were overfitting the data and removed the said features through feature engineering
3. After the removal, we obtained accuracy and R^2 of 0.820119583
4. and 0.714489414 respectively. And RidgeRegression of 0.795877039 and 0.754899399 for accuracy and R^2.
5. We finally used the Xgboost Regressor and obtained an accuracy of 0.944345864 and an R^2 score of 0.941036641.

Result

We used the Xgboost model to build the application and used streamlit for the UI. We avoided variables like Country, Status and HIV that would cause the model to overlearn these features. Thus we obtained a model with reasonably high accuracy without compromising the intent of the competition.