Computational Machine Learning (COSC2793)

Assignment 1: Introduction to Machine Learning

Vishwa Gandhi - S3714805

Contents

Introduction

The aim of the project is to predict the life expectancy of new born based on the some information given about regions by using appropriate approach. Major steps implemented to explore the data, approaches used to predict the result and evaluation techniques used to make the judgment. Assumptions made and analogy used is discussed.

The data provided is from "Global Health Observatory data repository". It is already cleaned. Hence, only changes made to data are to improve the efficiency of approach taken. Data has continuous variables along with some categorical numerical variables.

All the parameters selected for the models are manually engineered before selecting them in order to achieve the best result.

Implementation

This section explains steps taken in order to achieve the goal. This includes data loading, data exploration, data engineering and analysis. This assignment is created in python using libraries like pandas, matplotlib, numpy, sklearn, seaborn. The data is loaded into DataFrame for further processing.

Data Exploration and Data Engineering

- After loading the data, statistical analysis suggests that features have very different range of the values.

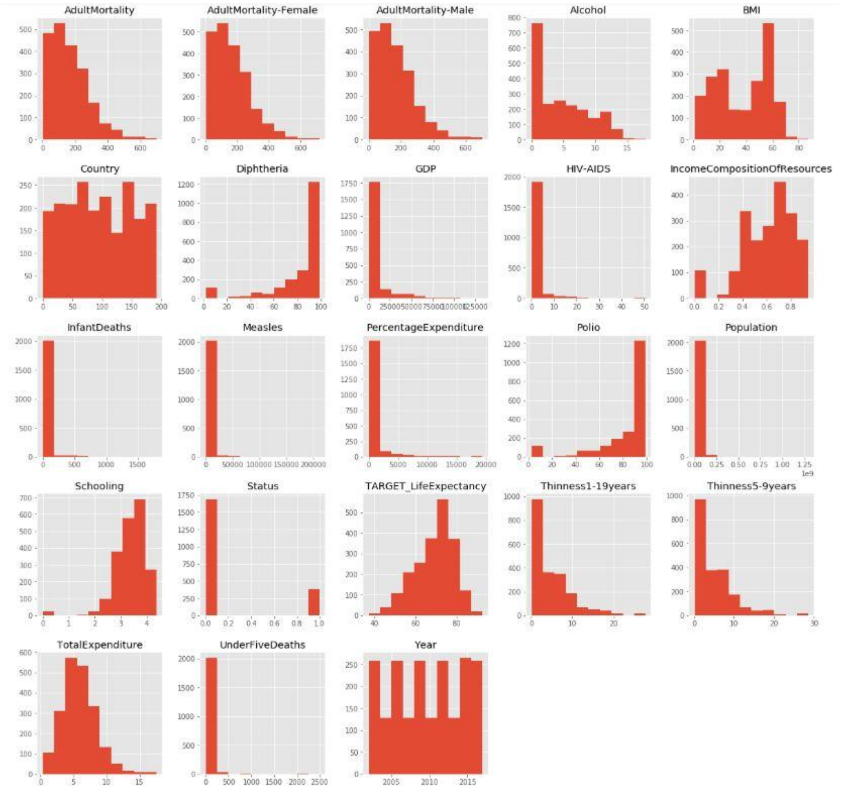

Figure 1: Distribution of descriptive features

Histograms of features in the figure indicate that few features has very broad range of value and some having a small range. In order to normalize the data to avoid the effect of the scale on algorithm performance, MinMaxScaler() is applied on all the variables excluding target variable. This eventually converts all the values between 0 to 1 reducing the scale effect.

- Studying correlation between features help making decision of choosing correct approach. It is a fundamental property to analyze. Following heat map indicates that some features are considerably correlated to each other.

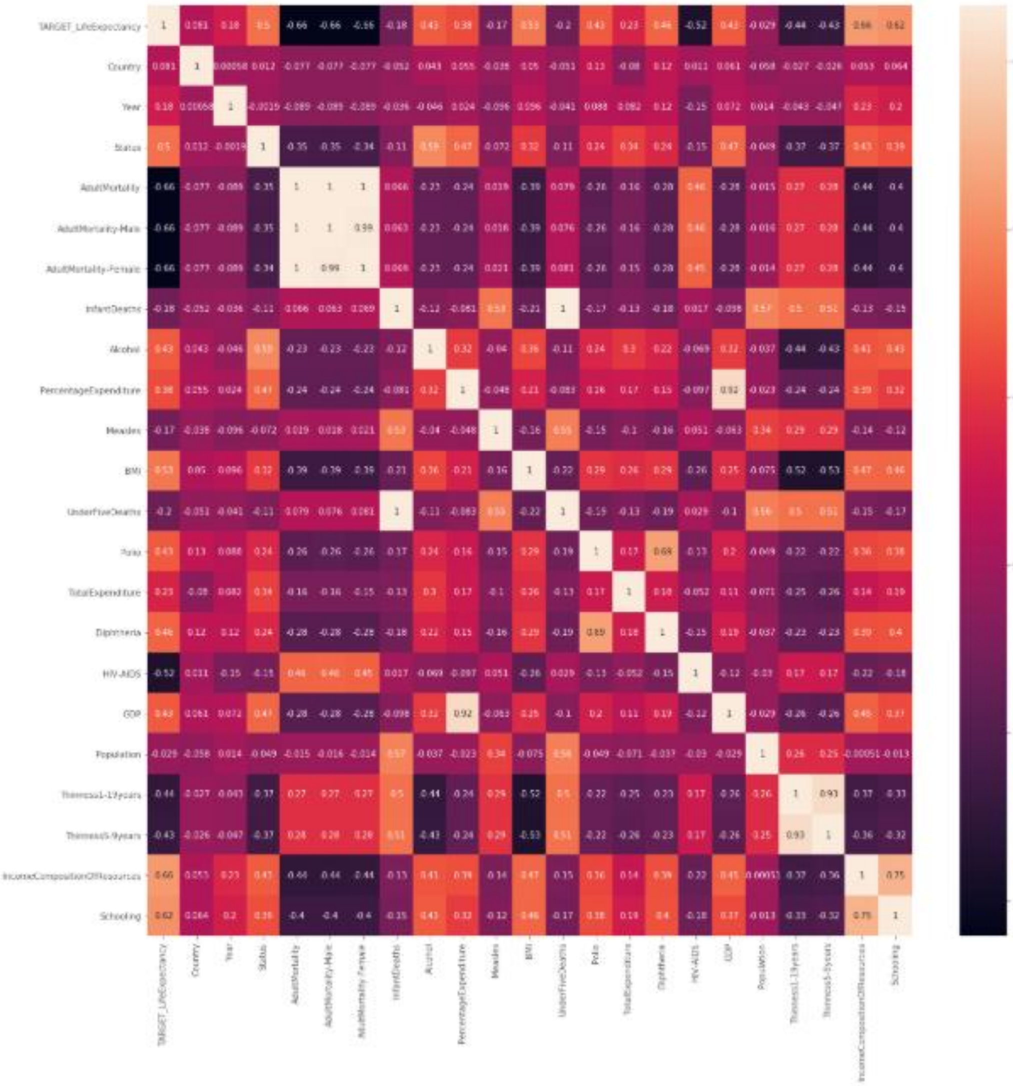

Figure 2: Heat Map of features for correlation analysis

Model Evaluation Strategy

- The data is divided in train and test set to for descriptive features and target variable. After detailed consideration of effect of split ratio on performance, data is divided in 80:20 train:test ratio to conduct further analysis. Note that this data is randomly divided.
- In order to evaluate the efficiency of algorithm, MSE and RMSE are used which are average model prediction errors. As RMSE is a square root of MSE, it assigns relatively high weights to large errors. Comparing the result in context of both gives better understanding about errors reducing performance. Additionally, R2 score is also used to evaluate if model explaining all the variability of the data around its mean.
- I have implemented various regression algorithms on the given data. This report has mention about few of them. To find out about other techniques, please refer the implementation code.

Major findings

- The first technique I have used is multiple linear regression. The data has multiple descriptive features and single target feature, so this makes multiple linear regression suitable in our case. This algorithm has MSE of 19.54.
- Polynomial regression is an extended form of linear regression which takes non linear relationship of variables into account. So I have implemented this algorithm to check the performance in multi dimensional space. This approach with degree 2 gives its best MSE of 23..46.
- Another method evaluated is Lasso Regression which uses shrinkage where values are shrunk towards central data point. Lasso eliminates least important features of the model by configuring very low value of certain coefficients. Poly Nomial Lasso Regression internally performs regularization. It has slightly better performance than above discussed 2 approaches with MSE of 16.208 at alpha value of 0.01.
- At the last, I have implemented Ridge regression with polynomial of 2. This approach adds a penalty to the update. This will result in shrinkage of coefficients. Ridge regression also performs regularization to balance bias and variance of the data internally. The major thing making ridge more suitable for this assignment is its ability to deal with multicollinearity. As we have seen in figure 2, there is correlation between predictor variables. For the data provided, Ridge Polynomial regression matched the best and has MSE of only 14.65 considering average case of all data splits.

Final Outcome

Detailed analysis of performance of regression approaches on the provided data has revealed below mentioned points:

- Ridge PolyNomial Regression approach with alpha of 1, is performing best on the given data case.

- Multiple variables exhibit correlation among them, which makes ridge suitable to be able handle that without affecting performance.
- Bias and Variance causes under-fitting and over-fitting of the model. Ridge has capability to handle that by shrinking the coefficient.
- Ridge technique does not drop any feature by driving the coefficient to 0 value like lasso regression. Hence it meets the requirement of the assignment of keeping all the features.

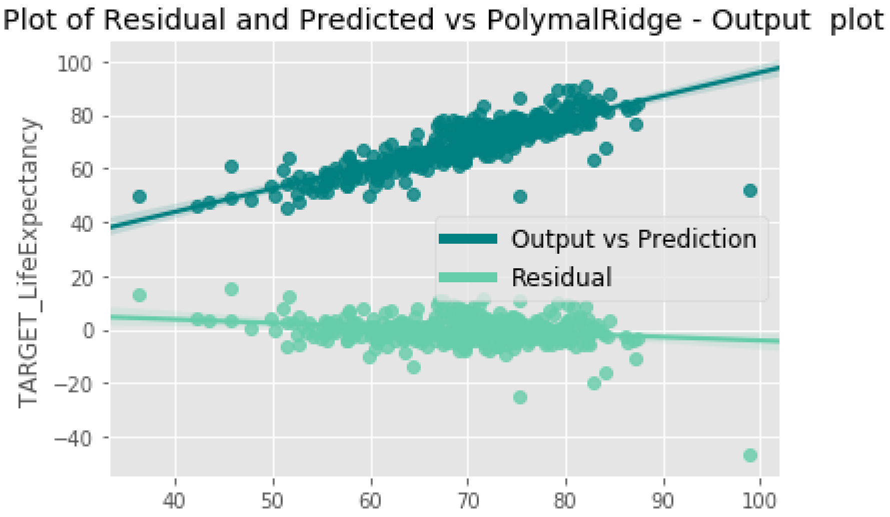

Figure 3: Plot of residual and Predicted vs PolyNomialRidge - output

This plot indicates the spread of residual and gap between output and predicted value for Ride Poly Nomial Analysis.

References

1. Apps.who.int. 2020. *GHO | By Category | Life Expectancy And Healthy Life Expectancy - Data By Country*. [online] Available at: <https://apps.who.int/gho/data/view.main.SDG2016LEXv?lang=en> [Accessed 5 April 2020].

2. Medium. 2020. *Learning Python Regression Analysis — Part 9: Tests And Validity For Regression Models*. [online] Available at: <https://medium.com/@dhwajraj/learning-python-regression-analysis-part-9-tests-and-validity-for-regression-models-78dcd5cde3a1> [Accessed 5 April 2020].

3. Medium. 2020. *A Beginner'S Guide To Linear Regression In Python With Scikit-Learn*. [online] Available at: <https://towardsdatascience.com/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-83a8f7ae2b4f> [Accessed 5 April 2020].

4. Ml-cheatsheet.readthedocs.io. 2020. *Linear Regression — ML Glossary Documentation*. [online] Available at: <https://ml-cheatsheet.readthedocs.io/en/latest/linear_regression.html> [Accessed 5 April 2020].