COSC2673 – Assignment 1

Student Details:-

- Vikas Virani : **s3715555**

Project:-

- Predict the life expectancy of a newborn based on several attributes (features) related to the region which he/she was born in using a **regression approach**.

Approach:-

The dataset was already given & processed, so there was no need to preprocess it. I've also checked for missing values in the data. After loading the dataset, I've plotted a box plot for each of the continuous attribute in the dataset to check for the distribution of values in each attribute. The box plots for each is as per the figure-1 in Appendix. As can be seen from the graph, many of the attributes have outliers. Since we are not supposed to handle outliers any differently, I've kept it as it is & the respective algorithm will take care of it differently. These variables are not visualized in other way as the changes in those are not allowed & has to work on it as it is. **Scaling** is used to scale the data (reduce the effect of outlier as well).

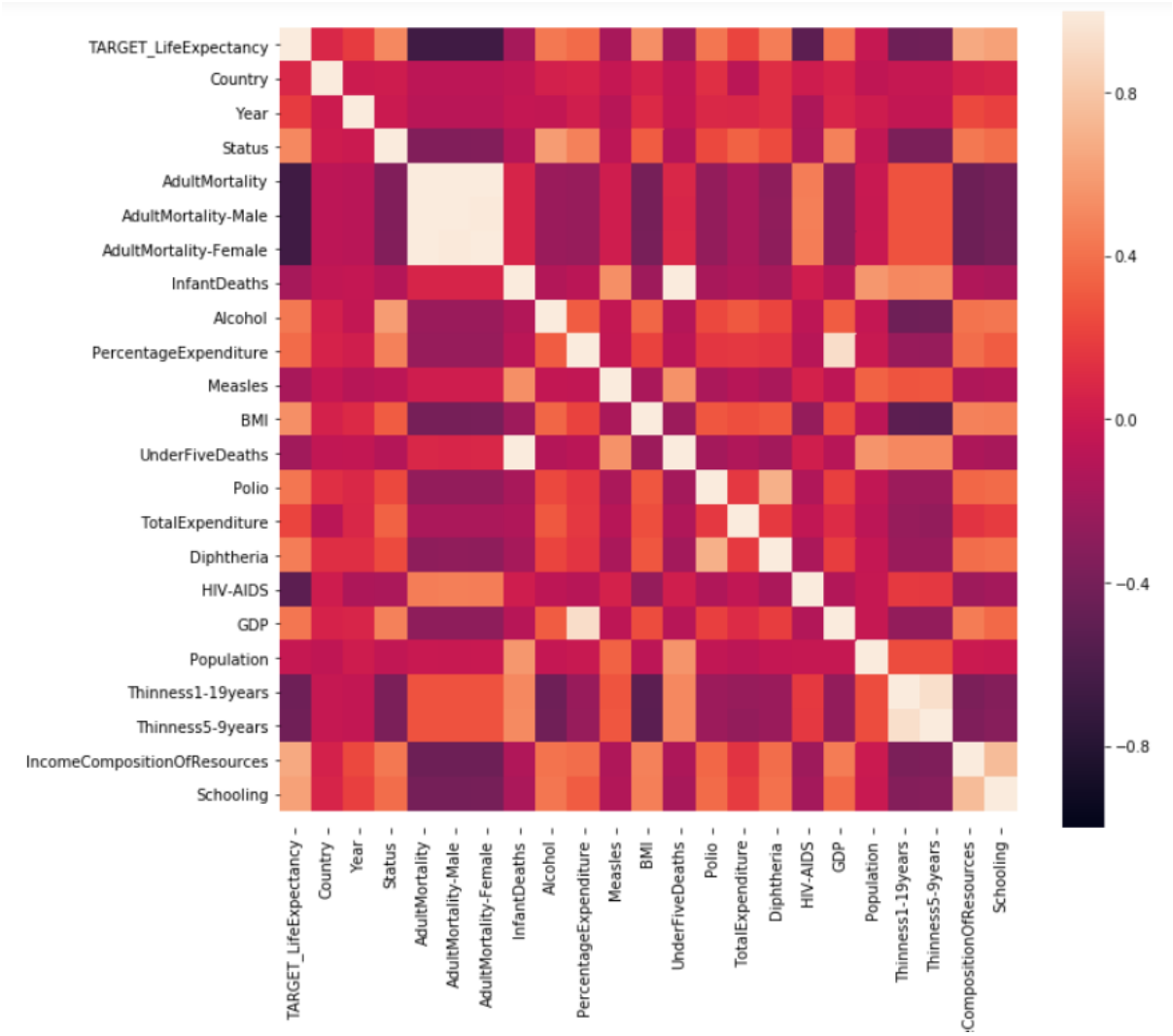
Essentially, below mentioned Regression techniques were explored in the analysis of this data:-

- Linear Regression
- Polynomial Regression
- Ridge/Lasso/ElasticNet Regression (Regularization technique)
- Ridge with Polynomial Regression
- TheilSen Regressor
- Bayesian ARD(Automatic Relevance Determination) regression

Out of which, Lasso/ElasticNet regularization technique & TheilSen Regressor and ARD regression are not included in the coding part as only best 3 algorithms in terms of Evaluation performance are added in the Jupyter file.

Summary statistics of all variables is explored and then heatmap is plotted to check multicollinearity of attributes. It is obvious from the below plot that "UnderFiveDeaths" & "InfantDeaths" are highly collinear. Also, "GDP" & "PercentageExpenditure" are collinear.

(Decision): If your linear model contains many predictor variables or if these variables are correlated, the standard OLS parameter estimates will have large variance, thus making the model unreliable. To overcome this, you can use regularization technique which allows to decrease this variance at the cost of introducing some bias. Finding a good bias-variance trade-off allows to minimize the model's total error. This was the rationale behind using **regularization technique** along with regression model.



We can use either of the regularization technique from Ridge, Lasso & ElasticNet. Of course, Lasso can set some coefficients to zero, thus performing variable selection (i.e. **Feature Selection**), while ridge regression cannot. But, Lasso tends to do well if there are a small number of significant parameters and the others are close to zero (i.e. when **only a few predictors** actually influence the response) while Ridge works well if there are many large parameters of about the same value (i.e. when **most predictors impact the response**).

Also, Ridge estimators are indifferent to multiplicative scaling of the data; i.e. if two variables are multiplied by a constant, the coefficients of fit does not change for a given λ parameter, however, for Lasso, the fit is not independent of the scaling. So if there are collinear variables & you need to include each of them in prediction then Ridge is a better choice.

(Evaluation measure): Selecting an effective evaluation measure is a crucial part in the analysis of machine learning algorithm. Here are some of the evaluation measures which can be in regression models [I]:-

- MAE/MSE (Mean Absolute/Square Error), Median Absolute Error, r2_score etc.

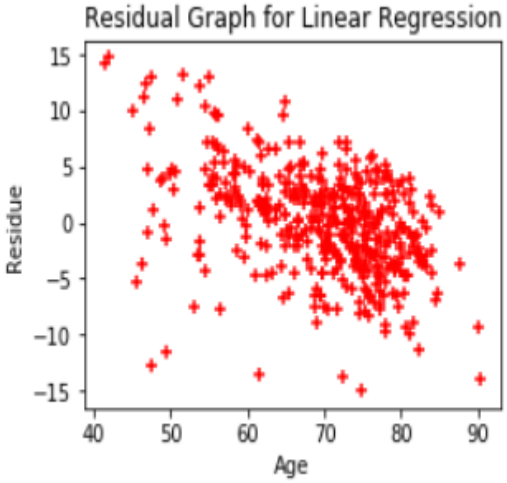
From the above measures, you can't rely on only one measure as it may not represent a true prediction power of machine earning algorithm. For example, R-squared only cannot determine whether the coefficient estimates and predictions are biased (It just represents how much variance in the target can be explained by our ML model), which is why you **must assess the residual plots**.

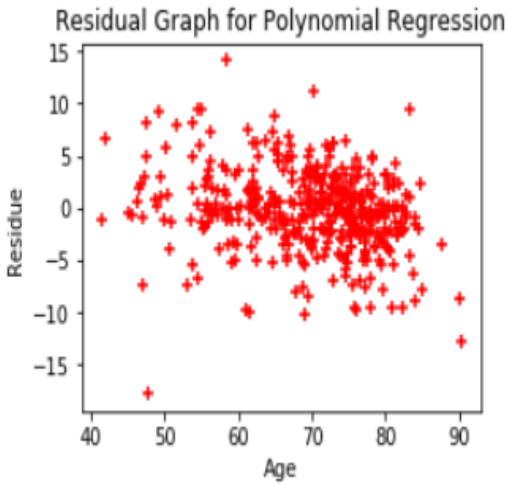
There are 2 problems associated with R-squared:-

- Every time you add a predictor to a model, the R-squared increases, even if due to chance alone. It never decreases. Consequently, a model with more terms may appear to have a better fit simply because it has more terms.
- If a model has too many predictors and higher order polynomials, it begins to model the random noise in the data. This condition is known as overfitting the model and it produces misleadingly high R-squared values and a lessened ability to make predictions

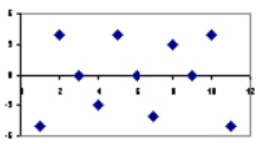
But R-squared values as part of evaluation are important to find out variance, so, I've decided to use both MSE & R-squared as evaluation measures & ultimate judgement about which model is best is based on the values of these two measures.

After Selecting Regularization technique & Evaluation measures, I've trained three models first; Linear Regression, Polynomial Regression & Ridge Regression [II]. As mentioned above, **residual graph** was plotted by me to check the validity of the model. Which are as below:-

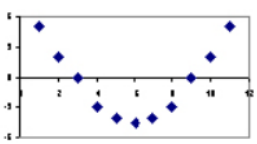




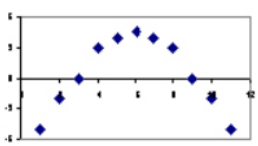
As per the data source from the references [IV] & [V] (Below graphs):



Random pattern



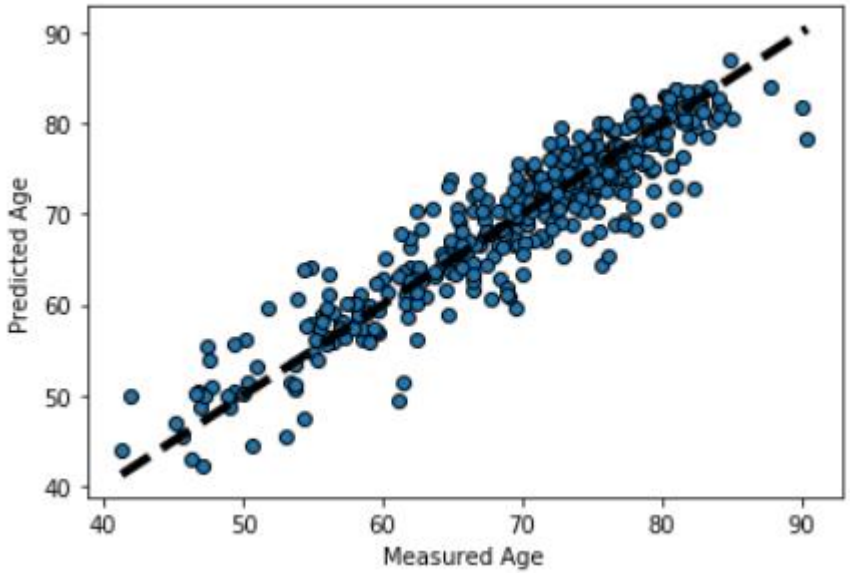
Non-random: U-shaped



Non-random: Inverted U

If your model have a random pattern in residual graph, then linear model can be a good fit for your data, but if your residual graph has a non-random pattern then it is a good fit for nonlinear model.

Out of these 3 model, Polynomial with Linear regression has the promising results among others. But as stated earlier, regularization is also needed in this dataset, so I took a model to one step further & build a **Polynomial with Ridge Regression** model & compared its evaluation measures with other ML models. Below scatter plot is the graph between Actual Age vs. Predicted Age by this model on test data:-



As can been seen from the graph, our last model has predicted the Age quite accurately, (i.e. All datapoints are near to our Line of Good fit). Also, below are the evaluation measures of each mode, as can be observed from these, our last model i.e. **Polynomial with Ridge Regression** is best for our data **& hence the ultimate judgement**. Also, it includes extra addition to simple linear model, like nonlinear/higher degree model as well as regularization which will reduce under fitting/Over fitting & maintains Bias/Variance tradeoff. Also, **GridSearchCV** is used to find the best parameter/hyperparameter tuning.

```
Mean squared error For Linear Regression: 22.530350559273952
R2 Score For Linear Regression: 0.7483888468572661
```

```
Mean squared error For Polynomial: 15.1090428646598
R2 Score For Polynomial: 0.8312674413095081
```

```
Mean squared error For Ridge: 22.5512973576007
R2 Score For Ridge: 0.7481549202670972
```

```
Mean squared error For Ridge with Polynomial: 12.961223660963297
R2 Score For Ridge with Polynomial: 0.8552535424206487
```

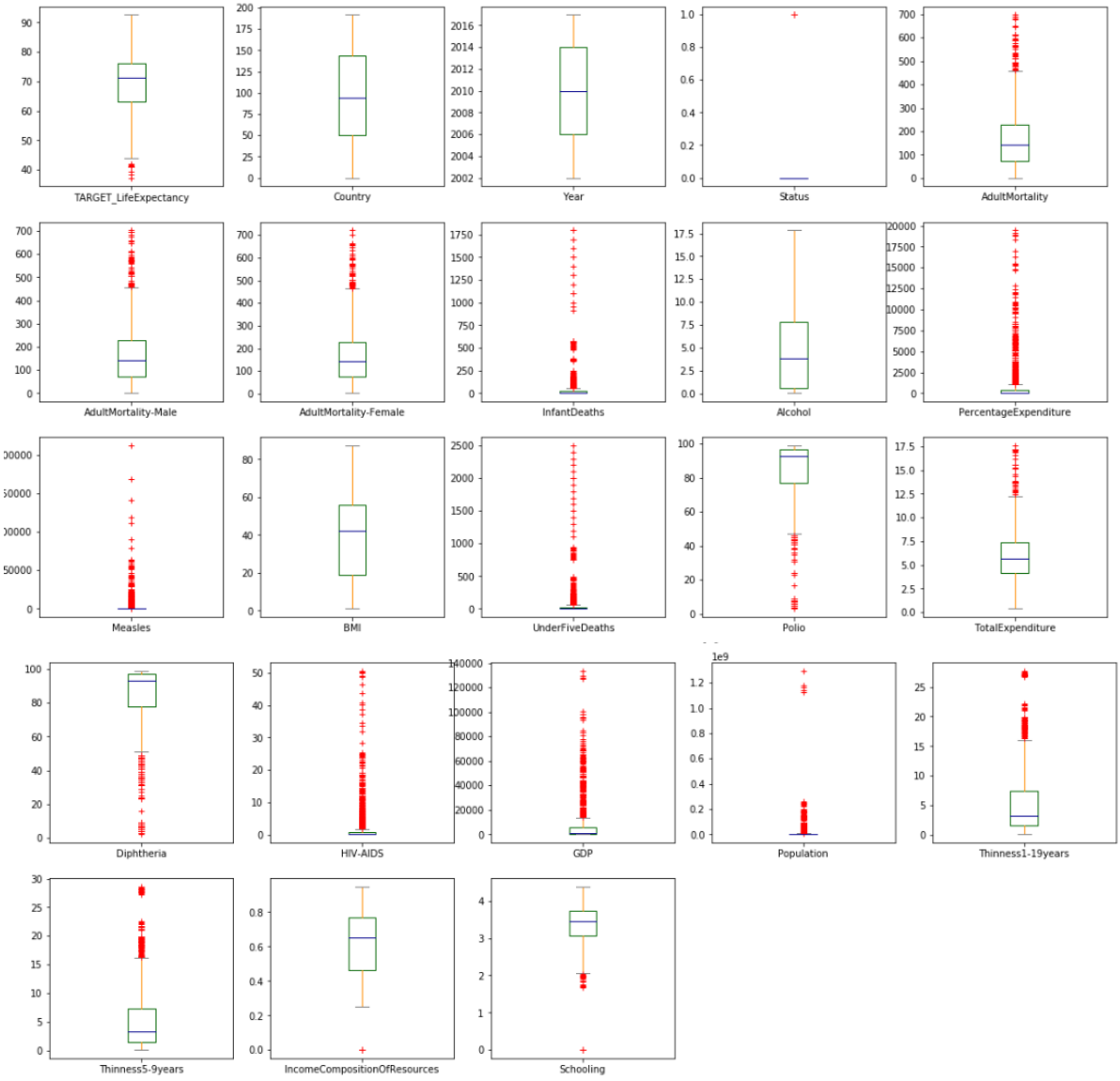
(We can also use **stats.ttest_rel(t-test)** from **scipy** or **F-1 test** to find the statistical significance difference between the evaluation measures of these model to find out the best statistically significant model among all)

References:-

I. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics
II. https://scikit-learn.org/stable/modules/linear_model.html#
III. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model
IV. https://blog.minitab.com/blog/adventures-in-statistics-2/linear-or-nonlinear-regression-that-is-the-question
V. https://stattrek.com/regression/residual-analysis.aspx
VI. https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-tutorial-and-examples
VII. https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
VIII. https://matplotlib.org/3.1.1/tutorials/introductory/pyplot.html

Appendix:-

I.


TARGET_LifeExpectancy
Country
Year
Status
AdultMortality
AdultMortality-Male
AdultMortality-Female
InfantDeaths
Alcohol
PercentageExpenditure
Measles
BMI
UnderFiveDeaths
Polio
TotalExpenditure
Diphtheria
HIV-AIDS
GDP
1e9
Population
Thinness1-19years
Thinness5-9years
IncomeCompositionOfResources
Schooling