RMIT UNIVERSITY

COSC2673 Semester 1 2024
Computational Machine Learning

Assignment 1
Introduction to Machine Learning

Nguyen Ngoc Khanh Linh

S3927588

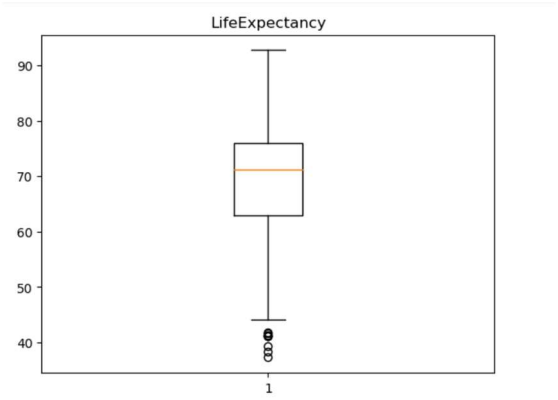
Contents

Exploratory Data Analysis

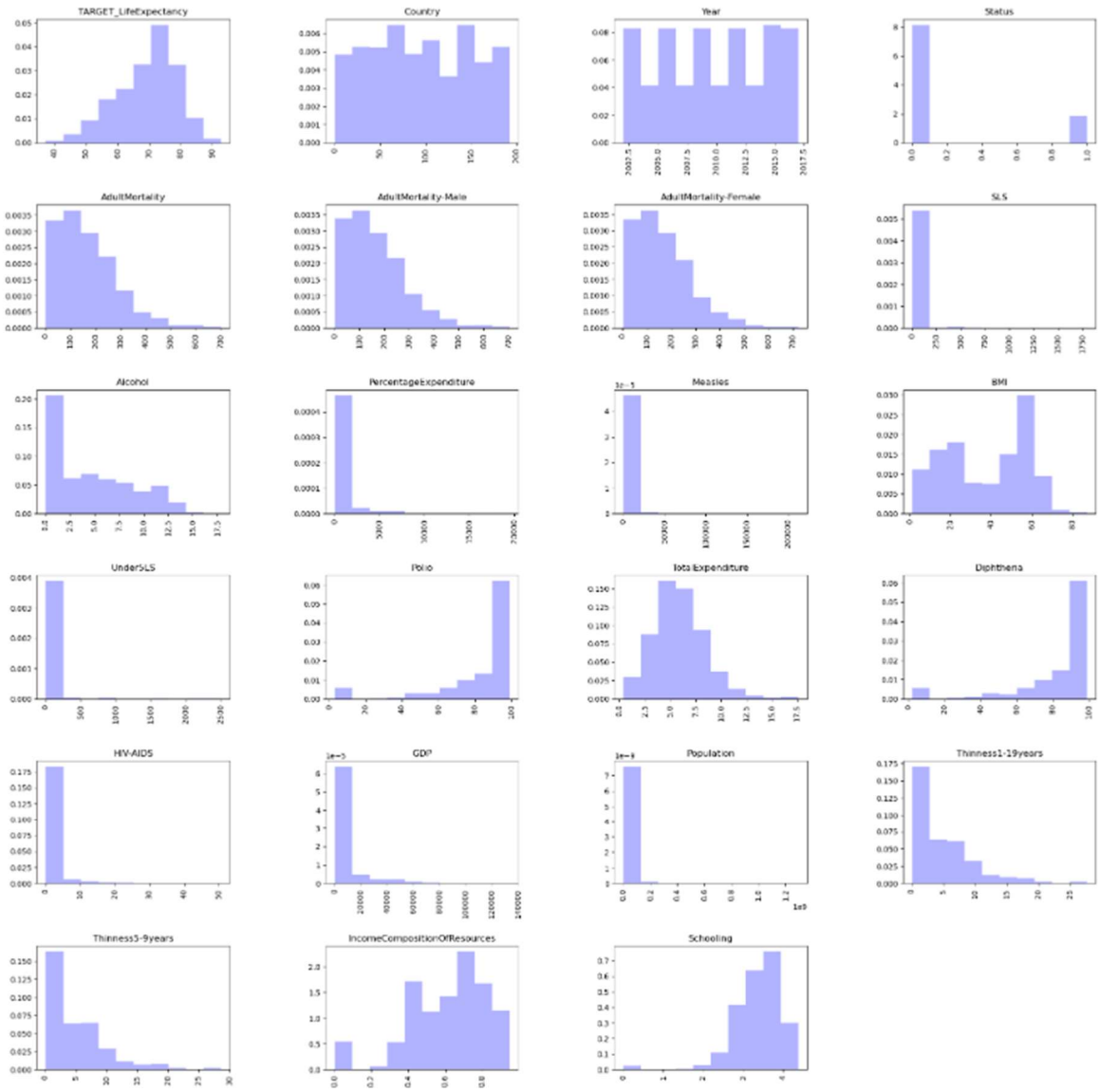
- Data frame's shape: 2071 entries and 23 columns
- Get summarised information of the dataset:

```
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   TARGET_LifeExpectancy           2071 non-null   float64
 1   Country                         2071 non-null   int64
 2   Year                            2071 non-null   int64
 3   Status                          2071 non-null   int64
 4   AdultMortality                  2071 non-null   int64
 5   AdultMortality-Male             2071 non-null   int64
 6   AdultMortality-Female           2071 non-null   int64
 7   SLS                             2071 non-null   int64
 8   Alcohol                         2071 non-null   float64
 9   PercentageExpenditure           2071 non-null   float64
 10  Measles                         2071 non-null   int64
 11  BMI                             2071 non-null   float64
 12  Under5LS                        2071 non-null   int64
 13  Polio                           2071 non-null   int64
 14  TotalExpenditure                2071 non-null   float64
 15  Diphtheria                      2071 non-null   float64
 16  HIV-AIDS                        2071 non-null   float64
 17  GDP                             2071 non-null   float64
 18  Population                      2071 non-null   int64
 19  Thinness1-19years               2071 non-null   float64
 20  Thinness5-9years                2071 non-null   float64
 21  IncomeCompositionOfResources    2071 non-null   float64
 22  Schooling                       2071 non-null   float64
dtypes: float64(12), int64(11)
```

  There are none missing values in the dataset. All entries are under float or integer.
- Numerical/ categorical variables:
    - Country and Status variables are categorical variables. Status is binary.
    - Numerical variables:
        - Continuous: Alcohol, Percentage Expenditure, BMI, Total Expenditure, Diphtheria, HIV-AIDS, GDP, Thinness1-19years, Thinness5-9years, Income Composition of Resources and Schooling.
- Box plot of TARGET_LifeExpectancy:

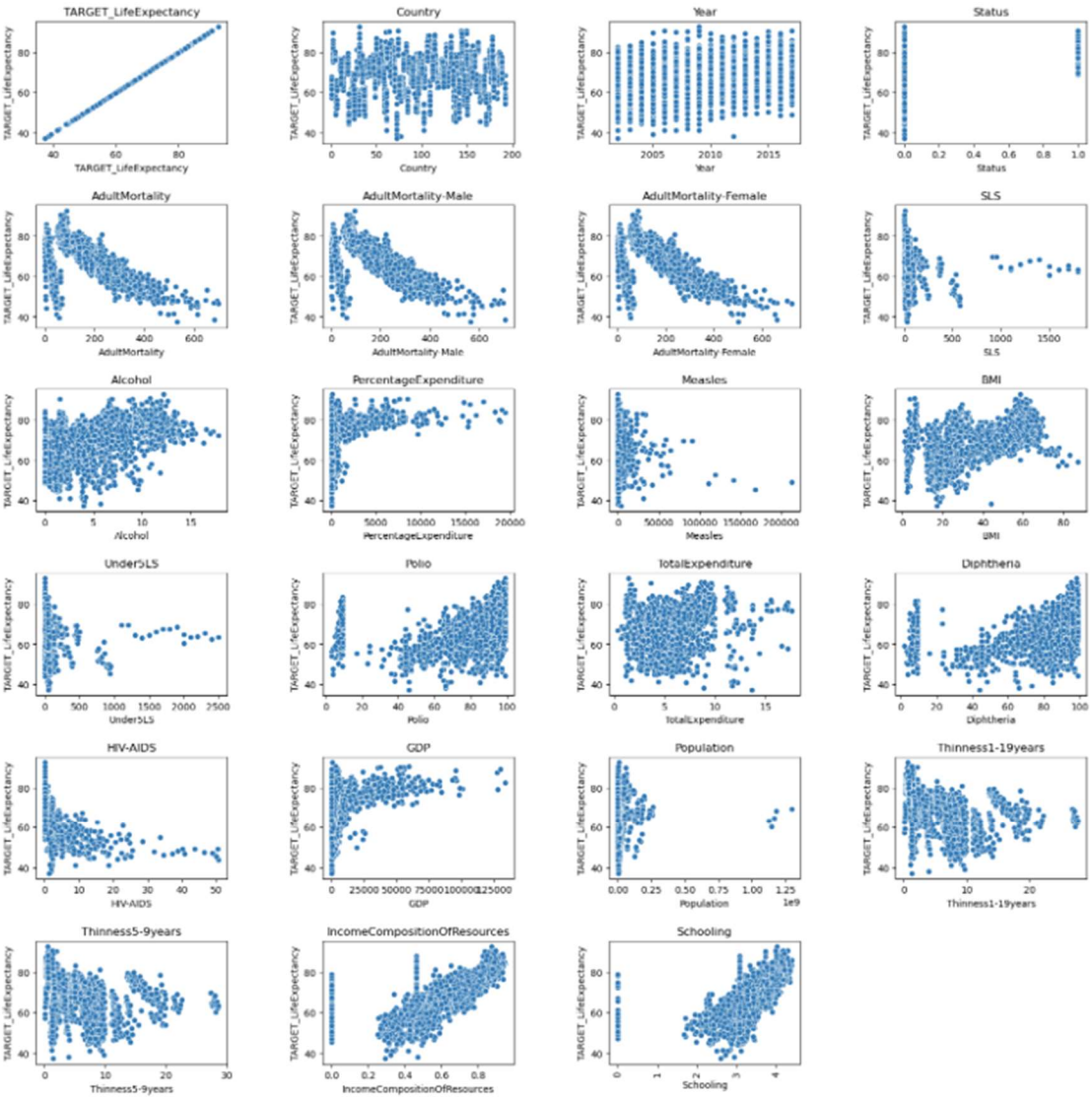


- Data distribution:

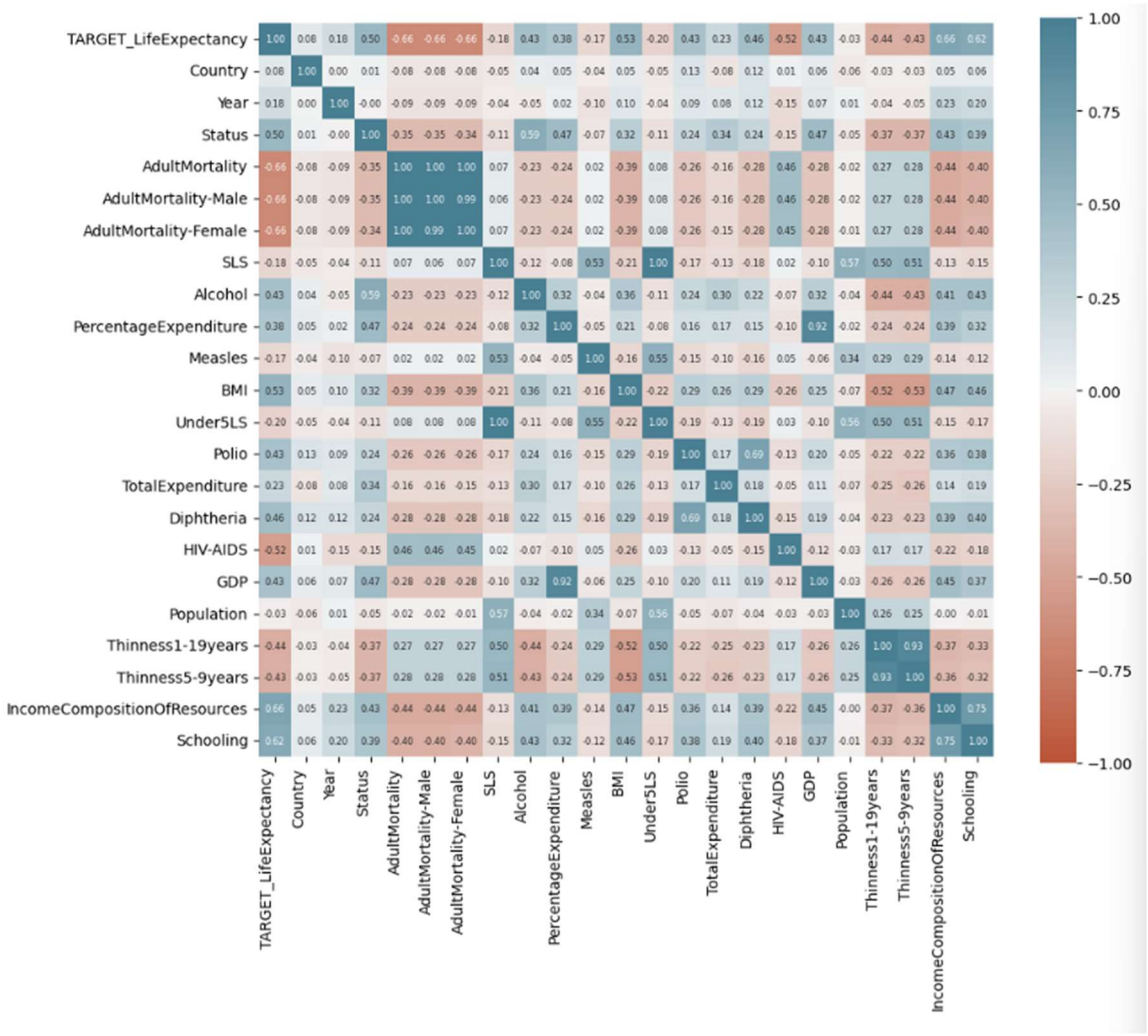


- Many attributes are heavily skewed including Adult Mortality, AdultMortalityMale, Adult Mortality Female, Alcohol, PercentageExpenditure, Measles, Under5LS, Polio, Diphtheria, HIV-AIDS, GDP, Population, Thinness1-19years, Thinness5-9years.
- Status is categorical attribute. Mostly data are categorized with class 0, and few instances with 1.
- Target variable Life Expectancy is distributed from 40 to 95, with some outliers falling below 45.

- Relationship between features and target variable



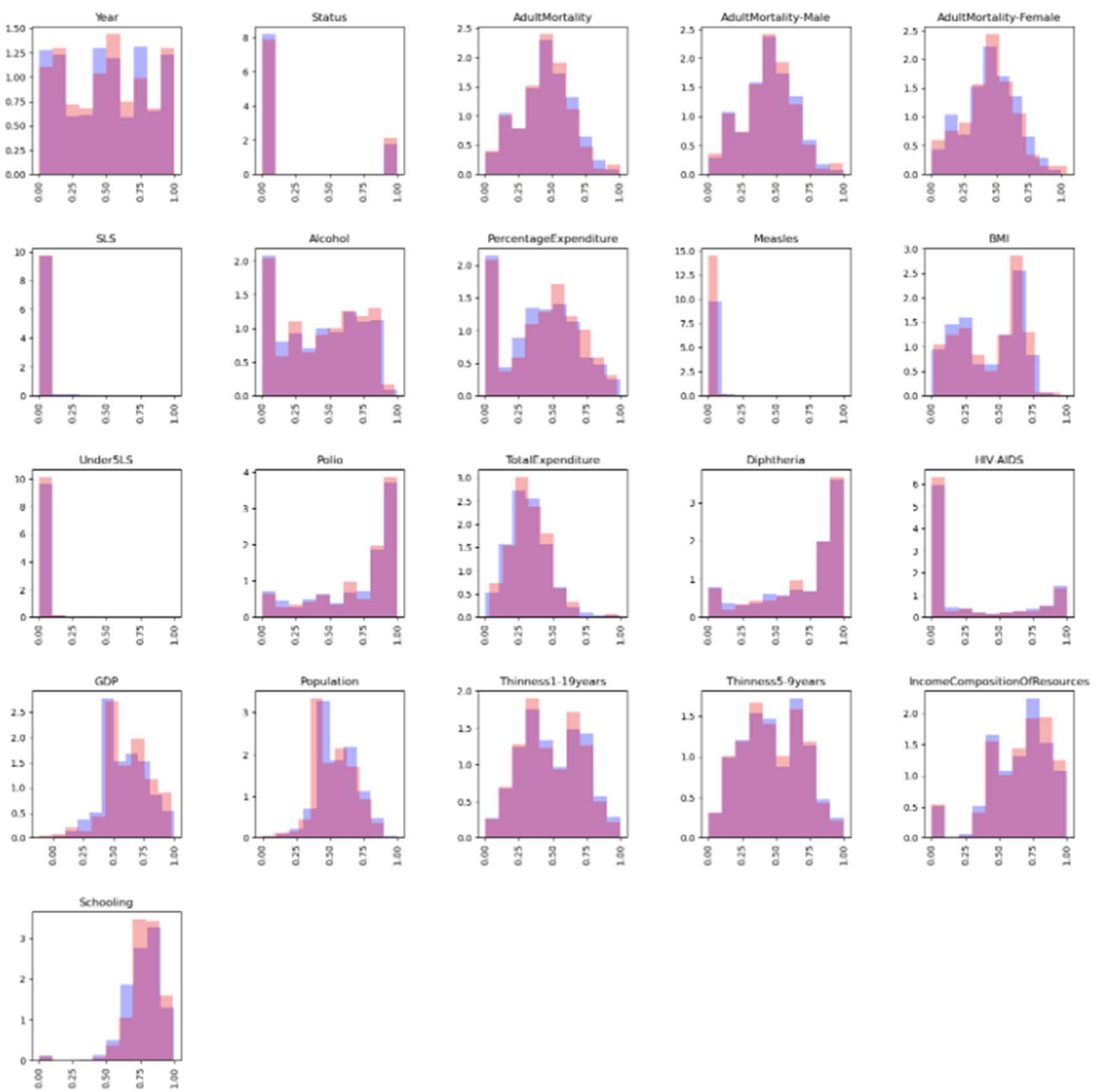
-
  - There seems to be a good linear relationship between LifeExpectancy and AdultMortality, IncomeCompositionOfResources and Schooling.
  - There are some attributes that seem to share nonlinear relationship with LifeExpectancy including Country, SLS, Measles, Year and Population.
- A heatmap with annotations supported in order to further clarify the correlation between variables.



-
-

Feature Scaling/ Normalisation

- Categorical attributes:
    - There are 2 categorical attributes including Country and Status. Since Status attributes has been under binary format (0 and 1), OneHotEncoder is only applied to encode the Country columns.
- Numerical attributes:
    - Log Normalisation is applied to transform highly skewed attributes including AdultMortality, Alcohol, PercentageExpenditure, Polio, Diphtheria, HIV-AIDS, Population, GDP, Thinness1-19years, and Thinness5-9years (witnessed from EDA). Applying Log Normalization for those highly skewed attributes made the distribution of the variable more symmetric and reduce the impact of extreme values.
    - Min-max scaling is also used to transform all numerical attributes.
- Check two splits are identically distributed after feature scaling and normalisation.



-

Baseline model and required improvements

- Choose **Linear Regression** to be baseline model since it is a simplest and easier linear model to train. The baseline model is trained on train dataset and then tested on the test set. Below are the loss scores for the Linear Regression:

```
Mean squared error of Linear Regression on training set: 7.527520905911634
Mean squared error of Linear Regression on testing set: 3.678705992896541e+23
Mean absolute error of Linear Regression on training set: 2.0824932301102055
Mean absolute error of Linear Regression on testing set: 33746448053.777138
R2 score of Linear Regression on training set: 0.9172658698666827
R2 score of Linear Regression on testing set: -4.3271835219197176e+21
```

It can be witnessed that the baseline model has a good score on train dataset. However, the trained model seems to not perform well on test set. There is a huge generalization gap between these 2 sets.

It suggested the model may be overfitting on the train dataset while not capturing the underlying patterns or relationships, leading to poor performance on unseen dataset.

Regularisation

In order to reduce the generalization gap, the other 2 linear models are trained: Ridge and Lasso. These 2 models are linear regression models that add a penalty term to the Ordinary Lease Squares (OLS) objective function. The penalty term added helps prevent overfitting.

- Ridge model performance: Below is the performance of Ridge model trained on train dataset and tested on the testing dataset.

```
Mean Squared Error (Training): 8.52042363146778
Mean Squared Error (Test): 9.329867263269467
Generalization Gap: -0.8094436318016864
```

  The default value for alpha in Ridge model here is 1.0.

- Lasso model performance: Below is the performance of Lasso model trained on train dataset and tested on the testing dataset.

```
Mean Squared Error (Lasso - Train): 43.540181020303464
Mean Squared Error (Lasso - Test): 39.93934801840146
Generalization Gap: 3.6008330019020036
Value of Alpha in Lasso regression: 1.0
```

  The default value for alpha in Lasso is also 1.0.

Discussion:

- It can be witnessed that these 2 models have successfully prevented overfitting in the regression model. However, it seems that the performance on the train set has been decreased lightly. The generation gap is narrowed down to a particularly great extent.

- In addition, comparation between performance of the 2 regularized models suggests that the trained Ridge model is performing better on the dataset compared to the other. This might be due to multicollinearity, where many attributes seem to have high correlations with the TARGET_LifeExpectancy.

Hyperparameter tuning

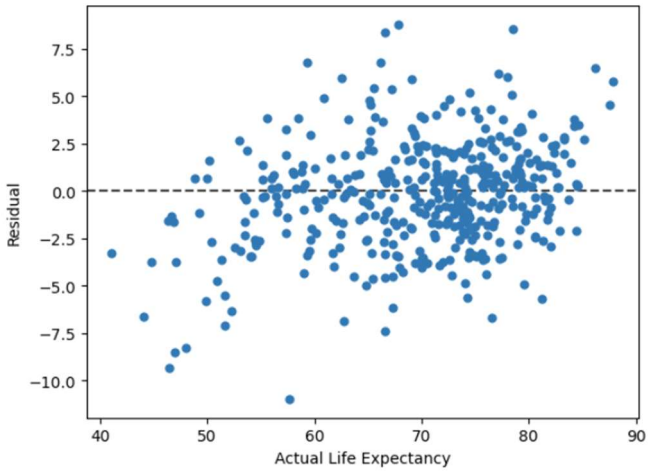
To find the best alpha values for the model, GridSearchCV is used to look for the best parameter. In here, the number of folds is set to 5 (**cv=5**). The value for scoring parameter will be **neg_mean_squared_error** measuring performance between the model and the data.

Tuning for Ridge:

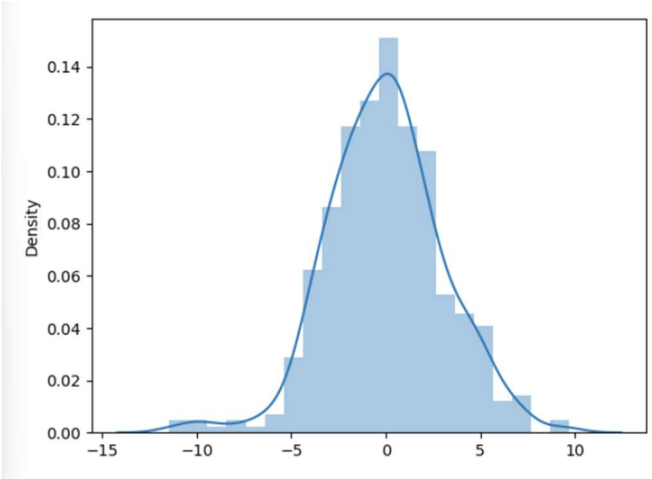
```
print('Best value for lambda:', ridgeRegressor.best_params_)
print('Best score for cost function:', ridgeRegressor.best_score_)
```

```
Best value for lambda: {'alpha': 0.001}
Best score for cost function: -9.898402213680539
```

Residual plot for Ridge final model:



Distribution plot

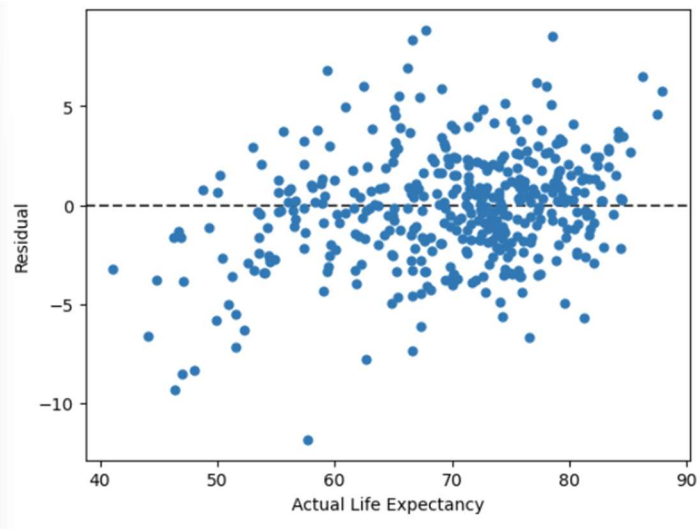


Tuning for Lasso:

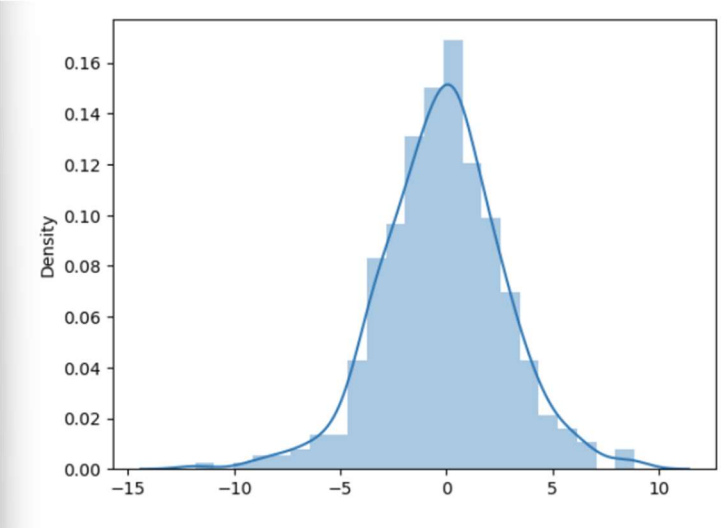
```
print("Best alpha value for Lasso:", LassoRegressor.best_params_)
print("Best score for the cost function:", LassoRegressor.best_score_)
```

```
Best alpha value for Lasso: {'alpha': 1e-15}
Best score for the cost function: -9.944072942180213
```

Residual plot for Lasso:



Distribution plot for Lasso model:



Evaluation methods

1. Cross validation:

- Instead of using a single train-test split, cross validation provides a more reliable result of the models' performance since it averages the performance over multiple folds.
- Reduce bias since each data points are used for both train and test set. Since the size of the dataset is small with just more than 2000 rows, cross validation is reasonable as all data points will be used.
- Prevent overfitting.

Here, all models are evaluated with the cross validation, the number of folds is 5. This is a reasonable number of folds compared to 10 as the size of the data set is not too large.

2. Mean Squared Error

- MSE is sensitive to errors with clear interpretability by presenting the average squared difference between predicted and actual values.

It can be seen from the statistics the Linear Regression, though possess the MSE for train set quite low, the MSE of the model performing on test set is far too high.

Meanwhile, comparing those 2 Ridge and Lasso, MSE from Ridge models seems to be lightly better compared to that of Lasso.

Ultimate judgement

- The ultimate model chosen for predicting TARGET_LifeExpectancy is the RidgeRegressor, as it has successfully prevented overfitting on the data. In addition, compared to Lasso which also added in the penalty term, the performance of Ridge is better when trained on the dataset.
- All models are evaluated based on Cross Validation and the main score to be assessed is Mean Square Error. However, as the MSE score is not too close to 0, which indicates there are still limitations in the model. These can be based on the pre-processing steps with data and the EDA. At the same time, hyperparameters tuning might be improved with better range of lambdas when further researched.
- A real-world settings in order to predict Life Expectancy may include the collection and analysis of data from various sources, including Demographic Information, Lifestyle factors, Medical History, Biometric Measurements, Geographical Factors and so on. Machine Learning models trained based on these data will also require a careful and thorough analysis. At the same time, many techniques will be required to improve and achieve high performance for the models built.