Innovaccer | IIT Madras | Decision Scientist-Test | Report

Name : Satish G
RollNo : CS14B042

* column = variable, variable = column

Data Preprocessing:

1. **Data Type change**
    - Some of the missing values in the data are represented by **'N/A'**(instead of NaN) which is a **string**, and as result of which the columns with have float values along with 'N/A' have their data type as **object** (non-numeric data type)
    - This will result in issue when handling non-numeric(categorical) data types. So, we have change those columns data type to a numeric data type(float64)

2. **Missing Values**
    a. **Target variable**('per_capita_exp_total_py')
        - Delete the rows in which target column value is missing
    b. **Numeric variables**
        - Replace the missing values in column with **mean** of the available values
    c. **Non-numeric variables**
        - Replace the missing values in column with **mode** of the available values
    d. **Majority missing**
        - If a column has missing values in majority(90%) rows, delete the column

3. **Majority variables**
    - Delete columns if majority(90%) rows have same value as the model won't learn anything from a feature in which change isn't much.
4. **Categorical variables**
    - Delete categorical variables as there are too many levels in each variable
    - Alternative is to cluster them(levels) based on their frequency -- didn't implement it
5. **Splitting**
    - Divide the data into ~85% train , ~7% valid, ~8% train data

Model building:

Linear Regression:

1. **No regularization**
    - Without regularization the model exactly(nearly) fits the data as the available training data is low
2. **Regularization(L2 -- Ridge)**
    - Use high values of alpha to avoid overfitting, since we have very low data
    - Optimal alpha -- alpha for which the model performs better on valid data
    - Even though in some of the data split scenarios the optimal model performance on test data is low, it's very high when compared to the no regularization model.