Assignment – Regression Algorithm

1. Identify your problem statement:

Stage 1: Machine Learning(inputs are present in numerical)

Stage 2: Supervised Learning(input and outputs are clear)

Stage 3: Supervised - Regression(inputs are numeric values)

2. Tell the basic info about your dataset:

There are 1338 rows × 6 columns in the given dataset.

Input columns/dependent: 'age', 'bmi', 'children', 'sex’, 'smoker’.

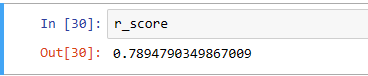
Output column/independent: 'charges'.

3. Mention the pre-processing method:

From the given dataset sex(male/female) and smoker(yes/no) columns are string format,so converting the 2 columns to nominal data(one hot encoding), as we can’t compare the data.

4. Develop a good model with r2\_score values:

* **MULIPLE LINEAR EGRESSION: (r value) = 0.7894**



* **SUPPORT VECTOR MACHINE:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| s.no | Hyper parameter | LINEAR  R value | RBF(NON LINEAR)  R value | POLY  R value | SIGMOID  R value |
| 1 | C10 | 0.4624 | -0.0322 | 0.0387 | 0.0393 |
| 2 | C100 | 0.6288 | 0.3200 | 0.6179 | 0.5276 |
| 3 | C500 | 0.7631 | 0.6642 | 0.8263 | 0.4446 |
| 4 | C1000 | 0.7649 | 0.8102 | 0.8566 | 0.2874 |
| 5 | C2000 | 0.7440 | 0.8547 | 0.8605 | -0.5939 |
| 5 | C3000 | 0.7414 | 0.8663 | 0.8598 | -2.1244 |

In SVM, R value **0.8663** is high in rbf(non linear), when hyperparameter, c=3000

* **DECISION TREE**

|  |  |  |  |
| --- | --- | --- | --- |
| S.No | **criterion** | **splitter** | R value |
| 1 | squared\_error | best | 0.6817 |
| 2 | squared\_error | random | 0.7207 |
| 3 | friedman\_mse | best | 0.6855 |
| 4 | friedman\_mse | random | 0.6954 |
| 5 | absolute\_error | best | 0.6576 |
| 6 | absolute\_error | random | 0.7246 |
| 7 | poisson | best | 0.6822 |
| 8 | poisson | random | 0.6077 |

In Decision Tree, R value **0.7246** is high in absolute\_error criterion and splitter is random.

* **RANDOM FOREST**

|  |  |  |  |
| --- | --- | --- | --- |
| ­S.No | criterion | n\_estimators and Random\_state = 0 | R value |
| 1 | squared\_error | 10 | 0.8331 |
| 2 | squared\_error | 100 | 0.8539 |
| 3 | friedman\_mse | 10 | 0.8332 |
| 4 | friedman\_mse | 100 | 0.8540 |
| 5 | absolute\_error | 10 | 0.8355 |
| 6 | absolute\_error | 100 | 0.8521 |
| 7 | poisson | 10 | 0.8178 |
| 8 | poisson | 100 | 0.8332 |

In Random Foest, R value **0.8540** is high in absolute\_error criterion and n\_estimators and Random\_state = 0.

6. Mention the final model and justiy:

The final model I have chosen is SUPPORT VECTOR MACHINE, since the r2\_score value **0.8663 which is nearly to 1**.