Assignment-Regression Algorithm

Problem Statement or Requirement:

A client’s requirement is, he wants to predict the insurance charges based on

the several parameters. The Client has provided the dataset of the same.

As a data scientist, you must develop a model which will predict the insurance

charges.

1. Identify your problem statement:
2. Supervised Learning:

Input and Output should be present and Client’s requirement should be very clear.

1. Machine Learning:

Client has provided the dataset in names and numbers

1. Regression:

Output is a numerical values so we should say Regression

1. Tell basic info about the dataset (Total number of rows, columns):

Total number of rows is 1338

Total number of columns is 6

1. Mention the pre-processing method if you’re doing any (like converting string to number nominal data) :

Nominal data used Algorithm is One Hot Encoding and Column can be Expanded

"sex" column, we can use one-hot encoding convert into two binary columns: "male" and "female." This encoding method represented as binary values (0 or 1)

"smoker" column, we can also use one-hot encoding, convert into two binary columns: "smoker-yes" and "smoker-no." This encoding method represented as binary values (0 or 1)

1. Develop a good model with r2\_score. You can use any machine learning algorithm; you can create many models. Finally, you have to come up with final model.

1.Multiple Linear Regression:

r2 value is 0.789479034986700

2.Support Vector Machine:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No | Hyper parameter | Linear  r2 value | Poly  r2 value | rbf  r2 value | Sigmoid  r2 value |
| 1 | C=10 | 0.566 | 0.159 | -0.018 | 0.073 |
| 2 | C=100 | 0.635 | 0.750 | 0.390 | 0.527 |
| 3 | C=500 | 0.765 | 0.859 | 0.696 | 0.490 |
| 4 | C=1000 | 0.744 | 0.860 | 0.828 | 0.143 |
| 5 | C=2000 | 0.741 | 0.860 | 0.860 | -2.584 |
| 6 | C=3000 | 0.741 | 0.860 | 0.868 | -6.826 |

SVM Regression r2 value is 0.868 using Hyper parameter is [C=3000, kernel = ”rbf” ]

3.Decision Tree:

|  |  |  |  |
| --- | --- | --- | --- |
| S.No | Criterion | Splitter | r2 value |
| 1 | Squared\_error | best | 0.690 |
| 2 | Squared\_error | random | 0.711 |
| 3 | Friedman\_mse | best | 0.685 |
| 4 | Friedman\_mse | random | 0.738 |
| 5 | Absolute \_error | best | 0.667 |
| 6 | Absolute\_error | random | 0.817 |
| 7 | Poisson | best | 0.730 |
| 8 | Poisson | random | 0.726 |

Decision Tree Regression r2 value is 0.817 using (Criterion = ”Absolute error”, Splitter = “random”)

1. Random Forest:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Criterion | Max\_Features | N\_Estimator | r2 value |
| 1 | Squared\_error | sqrt | 10 | 0.856 |
| 2 | Squared\_error | sqrt | 100 | 0.869 |
| 3 | Squared\_error | Log2 | 10 | 0.856 |
| 4 | Squared\_error | Log2 | 100 | 0.869 |
| 5 | Absolute\_error | sqrt | 10 | 0.841 |
| 6 | Absolute\_error | sqrt | 100 | 0.868 |
| 7 | Absolute\_error | Log2 | 10 | 0.841 |
| 8 | Absolute\_error | Log2 | 100 | 0.868 |
| 9 | Friedman\_mse | sqrt | 10 | 0.861 |
| 10 | Friedman\_mse | sqrt | 100 | 0.868 |
| 11 | Friedman\_mse | Log2 | 10 | 0.861 |
| 12 | Friedman\_mse | Log2 | 100 | 0.868 |
| 13 | Poisson | sqrt | 10 | 0.853 |
| 14 | Poisson | sqrt | 100 | 0.868 |
| 15 | Poisson | Log2 | 10 | 0.853 |
| 16 | Poisson | Log2 | 100 | 0.868 |

The Random Forest Regression R2 value is (criterion = ”squared\_error”,max\_features = “log2”,n\_estimators = “100”,r2 value = 0.869)

5.) All the research values (r2\_score of the models) should be documented.

(You can make tabulation or screenshot of the results.)

| **Model Name** | **r2\_score** |
| --- | --- |
| Model 1 (Multiple Linear Regression) | 0.789 |
| Model 2 (Support Vector Machine) | 0.868 |
| Model 3 (Decision Tree:) | 0.817 |
| Model 4 (Random Forest) | 0.869 |

6.) Mention your final model, justify why u have chosen the same.

The final machine learning best method of Regression:

Random Forest Regression is the highest r2\_value = 0.869

Regression Assignment :

Git hub link in Multiple linear Regression: <https://github.com/Vidhyavino/Regression-Assignment/blob/ef1a2ffbd1b7f1d7bf2dacc0ebd759b5eabf8da7/Multiple%20Linear%20Regression%20Algorithm.ipynb>

Git hub link in SVMR:

<https://github.com/Vidhyavino/Regression-Assignment/blob/ef1a2ffbd1b7f1d7bf2dacc0ebd759b5eabf8da7/%20SVMR.ipynb>

Git hub link in Decision Tree:

<https://github.com/Vidhyavino/Regression-Assignment/blob/ef1a2ffbd1b7f1d7bf2dacc0ebd759b5eabf8da7/%20Decision_Tree.ipynb>

Git hub link in Random Forest:

<https://github.com/Vidhyavino/Regression-Assignment/blob/ef1a2ffbd1b7f1d7bf2dacc0ebd759b5eabf8da7/%20Random%20Forest.ipynb>

Git hub link in Deployment phase in Random Forest:

<https://github.com/Vidhyavino/Regression-Assignment/blob/ef1a2ffbd1b7f1d7bf2dacc0ebd759b5eabf8da7/%20Deployment_Phase_Random%20Forest.ipynb>