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

3.) Mention the pre-processing method if you're doing any (like converting string to number – nominal data)

4.) 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.

5.) All the research values (r2_score of the models) should be documented. (You can make tabulation or screenshot of the results.)

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

Solution:-

1.) Identify your problem statement

Stage-1- Machine Learning

Why its ML?

In this problem statement having numberical data then I use Ml

Stage-2- ML under Supervised Learning

Why its Sl?

I n this problem statement cleared input and output then I decide to supervised learning

Stage-3-Sl under Regression

Why its Regression?

Problem statement is numberical type then only select regression

Under Regression

Multiple linear regression
More then one input and one output.

Then this model r2 value is good save the model otherwise perform other regression algorithms likes

Support vector machine
Decision tree
Random forest
Perform the model and which one give r2 value is good then save the model.

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

Total number of rows, columns = (1338 rows × 6 columns)

1.Multiple Linear Regression:-

R2 value= 0.78651

This model is poor result then performs the one by one other algorithm in regression .

2.**Support Vector Machine:-**

S.no	Hyperparameter	Linear (r2 value)	Rbf(non-linear)r2value	Sigmoid (R2value)	Ploy (r2value)
1	C=100	0.5218	-0.1550	-0.12415	-0.1319512
2	C=1000	0.6188	-0.14956	-1.52178	-0.092372
3	C=10000	0.76152	-0.05511	-109.2006	0.3031
4	C=100.500	0.5219	0.15510	-0.12436	-0.1319
5	C=251.20	0.58715	-0.15954	-0.19109	-0.1288
6	C=2500	0.629010	-0.13622	-7.7449	-0.018195

Here Support Vector Machine regression use hyperparameter and linear kernal r2 value=0.76152 is poor so need a best model so I can do decision tree.

3.**Decision Tree:-**

S.no	Criterion	Splitter	Max_Features	R2_value
1	squared_error	Best	auto	0.69212
2	squared_error	best	sqrt	0.693917
3	squared_error	best	Log2	0.67637
4	squared_error	random	auto	0.69444
5	squared_error	random	sqrt	0.6024
6	squared_error	random	Log2	0.693319
7	friedman_mse	best	Auto	0.690352
8	friedman_mse	Best	Sqrt	0.7464
9	friedman_mse	best	Log2	0.675978
10	friedman_mse	Random	Auto	0.694034
11	friedman_mse	Random	Sqrt	0.68304
12	friedman_mse	random	Log2	0.631394
13	absolute_error	Best	Auto	0.661121
14	absolute_error	Best	Sqrt	0.71097
15	absolute_error	Best	Log2	0.44272
16	absolute_error	Random	Auto	0.77714
17	absolute_error	Random	Sqrt	0.68543
18	absolute_error	Random	Log2	0.66977
19	poisson	Best	Auto	0.67751
20	Poisson	Best	Sqrt	0.67302
21	Poisson	Best	Log2	0.62083

22	Poisson	Random	Auto	0.7234
23	Poisson	Random	Sqrt	0.7385
24	poisson	Random	Log2	0.57016

This Decision tree regression use criterion ,splitter and max_features (r2 _value)=0.77714
This value also poor.so need a best model so I can do Random forest .

4.Random Forest:-

S.no	Criterion	n_estimators	Max_Features	R2_value
1	squared_error	100	auto	0.85765
2	squared_error	200	sqrt	0.872083
3	squared_error	500	Log2	0.87222
4	friedman_mse	100	auto	0.8585
5	friedman_mse	200	sqrt	0.87376
6	friedman_mse	500	Log2	0.87119
7	absolute_error	100	Auto	0.85627
8	absolute_error	200	Sqrt	0.8732
9	absolute_error	500	Log2	0.874876
10	poisson	100	Auto	0.830606
11	poisson	200	Sqrt	0.83594

12	poisson	500	Log2	0.83625

This Random Forest regression use criterion ,n_estimators and max_features (r2 _value)=0.874876

Hance the above all regression algorithm are validating the data finally random forest regression is best among all regression algorithms