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

Machine Learning - Supervised - Regression

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

Rows- 1338

Columns – 6 (age, sex, bmi, children, smoker, charges)

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

Yes, converting string columns (sex and smoker) to number using One Hot Encoding method.

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.

Final model: Random Forest Algorithm

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

1. ***Multiple Linear Regression***:

R^2 Value= 0.7894

2. ***Support Vector Machine:***

R^2 Value= 0.7648 (linear, C=5000)

S.No	**Hyper Parameter**	**Linear (r^2 value)**	**Rbf (r^2 value)**	**Poly (r^2 value)**	**Sigmoid (r^2 value)**
1	C=10	-0.0016	-0.081	-0.093	-0.090
2	C=100	0.5432	-0.128	-0.097	-0.118
3	C=500	0.5902	-0.124	-0.082	-1.665
4	C=1000	0.6340	-0.117	-0.055	-5.616
5	C=2000	0.6893	-0.107	-0.002	-12
6	C=3000	0.7590	-0.096	0.0489	-12
7	C=5000	0.7648	-0.073	0.1462	-31
8	C=10000	0.7444	-0.017	0.3529	-119

3. ***Decision Tree:***

R^2 Value= 0.7338 (friedman_msc, log2, random)

S.No	**Criterion**	**Max_features**	**Splitter**	**r^2 value**
1	Squared_error	sqrt	best	0.7310
2	Squared_error	sqrt	random	0.6310

3	Squared_error	log2	best	0.6673
4	Squared_error	log2	random	0.7253
5	Friedman_msc	sqrt	best	0.7107
6	Friedman_msc	sqrt	random	0.6628
7	Friedman_msc	log2	best	0.7231
8	Friedman_msc	log2	random	0.7338
9	Absolute_error	sqrt	best	0.7002
10	Absolute_error	sqrt	random	0.5368
11	Absolute_error	log2	best	0.7157
12	Absolute_error	log2	random	0.6338
13	Poisson	sqrt	best	0.6657
14	Poisson	sqrt	random	0.6929
15	Poisson	log2	best	0.6838
16	Poisson	log2	random	0.6661

4. ***Random Forest***:

R^2 Value= 0.8741 (friedman_msc, log2, 100)

S.No	**Criterion**	**Max_features**	**N_estimators**	**r^2 value**
1	Squared_error	sqrt	10	0.8605
2	Squared_error	sqrt	100	0.8691
3	Squared_error	log2	10	0.8451
4	Squared_error	log2	100	0.8679
5	Friedman_msc	sqrt	10	0.8549
6	Friedman_msc	sqrt	100	0.8679
7	Friedman_msc	log2	10	0.8546
8	Friedman_msc	log2	100	0.8741
9	Absolute_error	sqrt	10	0.8550

10	Absolute_error	sqrt	100	0.8692
11	Absolute_error	log2	10	0.8341
12	Absolute_error	log2	100	0.8738
13	Poisson	sqrt	10	0.8458
14	Poisson	sqrt	100	0.8704
15	Poisson	log2	10	0.8541
16	Poisson	log2	100	0.8730

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

The model created in Random Forest Algorithm seems to be the best model as the r^2 value is nearer to 1 (For best model r^2 value ranges - 0 to 1) comparing to other r^2 value of other algorithms. The hyper parameters used in Random Forest Algorithms are criterion as friedman_msc, n_estimators as 100 and max_features as log2.

Random Forest - R^2 Value = 0.8741