

Assignment-Regression Algorithm

Problem Statement or Requirement:

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

- 1 Identify your problem statement
 - 1.Machine learning(input given number in datasheet)
 - 2.supervised learning(input and output clear)
 - 3.Regression(output given numerical)
- 2 Tell basic info about the dataset (Total number of rows, columns)
 - 1.Total number of rows-1338
 - 2.Total columns-6
- 3 Mention the pre-processing method if you're doing any (like converting string to number – nominal data)
dataset=pd.get_dummies(dataset,drop_first=True)- it is not show the dummies values
- 4 Develop a good model with r2_score. You can use any machine learning algorithm; you can
Algorithm: Support Vector Machine& Random Forest
Parameter: RBF,Criterion
max_features: SQRT
r2 value:0.874
- 5 All the research values (r2_score of the models) should be documented.
 - 1.MULITPLE LINEAR REGRESSION:
The Multiple Linear REGRESSION use R2 value=**0.7894**

2. SUPPORT VECTOR MACHINE:

Model.NO	HYPER PARAMETER	LINEAR value)	(r	RBF(NON LINEAR) value)	(r	Poly (r value)	SIGMOID (r value)
1	C10	0.462		-0.032		0.038	0.039
2	C100	0.628		0.32		0.617	0.527
3	C1000	0.764		0.81		0.617	0.287
4	C5000	0.741		0.874		0.859	-7.53

3. DECISION TREE:

Model.NO	criterion	max_features	splitter	r value
1	mse	auto	Random	0.741
2	mse	auto	best	0.711
3	mse	sqrt	Random	0.682
4	mse	sqrt	best	0.722
5	mse	log2	Random	0.659
6	mse	log2	best	0.741
7	mae	auto	Random	0.734
8	mae	auto	best	0.683
9	mae	sqrt	Random	0.647
10	mae	sqrt	best	0.732

11	mae	log2	Random	0.668
12	mae	log2	best	0.746
13	friedman_mse	auto	Random	0.707
14	friedman_mse	auto	best	0.669
15	friedman_mse	sqrt	Random	0.736
16	friedman_mse	sqrt	best	0.711
17	friedman_mse	log2	Random	0.765
18	friedman_mse	log2	best	0.735

4. Random Forest:

Model.NO	criterion	max_features	n_estimators	r value
1	mse	auto	10	0.843
2	mse	auto	50	0.858
3	mse	auto	100	0.852
4	mse	sqrt	10	0.855
5	mse	sqrt	50	0.868
6	mse	sqrt	100	0.874
7	mse	log2	10	0.861
8	mse	log2	50	0.871
9	mse	log2	100	0.871
10	mae	auto	10	0.845
11	mae	auto	50	0.856
12	mae	auto	100	0.851
13	mae	sqrt	10	0.864
14	mae	sqrt	50	0.867
15	mae	sqrt	100	0.872
16	mae	log2	10	0.853
17	mae	log2	50	0.871
18	mae	log2	100	0.871
19	friedman_mse	auto	10	0.838
20	friedman_mse	auto	50	0.849
21	friedman_mse	auto	100	0.849
22	friedman_mse	sqrt	10	0.856
23	friedman_mse	sqrt	50	0.867
24	friedman_mse	sqrt	100	0.871
25	friedman_mse	log2	10	0.854
26	friedman_mse	log2	50	0.868
27	friedman_mse	log2	100	0.868

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

Final model:

① The Random Forest Regression use R2 value(0.874),criterion=mse, max_features=sqrt, n_estimators=100

② The Support Vector Machine use R2 value(0.874), rbf(non-linear)

This two r2_value is good compared to another model