

Machine Learning - Regression Assignment

1. Identify your problem statement
- Goal : INSURANCE CHARGE PREDICTION
- 3 Stages : Machine Learning (Number based data)
- Supervised Learning (Input & Output present)
- Regression - MLR, SVM, DC, RF (Number based Output)
2. No of rows & columns : 1338 rows 6 columns
3. Pre-processing method : Categorical data - Nominal - One Hot Encoding Method (pandas function - get_dummies)

Multiple Linear Regression R2 score = 0.7978

Support Vector Machine :

r2 score				
S.No	Hyper Parameter	Linear	rbf (nonlinear)	poly (nonlinear)
1	C = 1	-0.012	-0.0987	-0.0899
2	C = 10	0.5007	-0.0423	0.0459
3	C = 100	0.6423	0.3547	0.6591
4	C = 500	0.7723	0.7094	0.8457
5	C = 1000	0.7501	0.8283	0.8631
6	C = 2000	0.7485	0.8625	0.8642
7	C = 3000	0.7485	0.8751	0.8632
8	C= 4000	0.7485	0.8803	0.8635
9	C= 5000	0.7484	0.8834	0.8629

The SVM regression algorithm R2 score (kernel = rbf (non-linear) & Hyper parameter (C = 5000)) = 0.8834

Decision Tree :

S.No	Criterion	Splitter	max_features	R2 score
1	squared_error	best	sqrt	0.7874
2	squared_error	best	log2	0.748
3	squared_error	random	sqrt	0.706
4	squared_error	random	log2	0.5437
5	friedman_mse	best	sqrt	0.7275
6	friedman_mse	best	log2	0.7686
7	friedman_mse	random	sqrt	0.7801
8	friedman_mse	random	log2	0.7826
9	absolute_error	best	sqrt	0.644
10	absolute_error	best	log2	0.7525
11	absolute_error	random	sqrt	0.6879
12	absolute_error	random	log2	0.7407
13	poisson	best	sqrt	0.7745
14	poisson	best	log2	0.791
15	poisson	random	sqrt	0.7072
16	poisson	random	log2	0.7365
17	Default parameter			0.7189

The Decision Tree regression algorithm R2 score (criterion = 'poisson', splitter = 'best', max_features = 'log2') = 0.7910

Random Forest:

S.No	n_estimators	Criterion	max_features	R2 score
1	50	squared_error	sqrt	0.8942
2	50	squared_error	log2	0.8915
3	100	squared_error	sqrt	0.8917
4	100	squared_error	log2	0.8947
5	1000	squared_error	sqrt	0.8944
6	1000	squared_error	log2	0.8951
7	50	friedman_mse	sqrt	0.891
8	50	friedman_mse	log2	0.8852
9	100	friedman_mse	sqrt	0.895
10	100	friedman_mse	log2	0.8918
11	1000	friedman_mse	sqrt	0.8947
12	1000	friedman_mse	log2	0.894
13	50	absolute_error	sqrt	0.8922
14	50	absolute_error	log2	0.8873
15	100	absolute_error	sqrt	0.8929
16	100	absolute_error	log2	0.8946
17	1000	absolute_error	sqrt	0.8947
18	1000	absolute_error	log2	0.8951
19	50	poisson	sqrt	0.8991
20	50	poisson	log2	0.8952
21	100	poisson	sqrt	0.8944
22	100	poisson	log2	0.8937
23	1000	poisson	sqrt	0.8942
24	1000	poisson	log2	0.8958
25	Default parameter	-	-	0.8743

The Random Forest regression algorithm R2 score (Almost all the parameters ranges between) = 0.87 to 0.89

Result :

2. R2_score values of all the algorithms mentioned is not above 0.90 (Good Model accuracy must be atleast above 0.95)
3. Based on the output I declare the dispersion of the given data is in Circular manner, since it is proved in SVM (rbf (non-linear) kernel)
4. So, we can move further on other Algorithms for better prediction with good accuracy.