HOPE AI

Assignment-Regression Algorithm - Dr. N. Dilip Raja

Problem Statement: Predict the insurance charges based on the given parameters

Basic information about the dataset: The dataset consists of six columns and 1337 rows. Out of this, five columns contain input data and one column shows the output values to train the ML model.

Pre-processing method: A preprocessing step is carried out to convert an input column containing string values. Since the column contains nominal data, we can use the **one-hot encoding** technique.

Model selection is carried out and the results from the several models are shown below

Multilinear Regression model

R^2 = 78.95%

Support Vector Machine Regression model

Kernal type	C value	R^2
Linear	0.01	-0.0888
	0.1	-0.0809
	1	-0.0101
	10	0.4624
	100	0.6288
	1000	0.7649
Poly	0.01	-0.0895
	0.1	-0.0883
	1	-0.0756
	10	0.0387
	100	0.6179
	1000	0.8566
72rbf	0.01	-0.089
	0.1	-0.089
	1	-0.0833
	10	-0.0322
	100	0.32
	1000	0.8102
Sigmoid	0.01	-0.0895
	0.1	-0.0882
	1	-0.0754
	10	0.0393
	100	0.5276
	1000	0.2874
Precomputed	0.01	Not computed because of technical problem

Decision Tree Regression model

Criterion	Splitter	R^2 (%)
Squared Error	Best	68.06
	Random	72.32
Friedman_mse	Best	70.12
	Random	76.32
Absolute Error	Best	68.97
	Random	72.76
Poisson	Best	71.82
	Random	69.82

Random Forest Regression model

Criterion	n_estimators	R^2 (%)
Squared Error	10	84.23
	50	85.61
	100	85.55
Absolute Error	10	84.17
	50	86.04
	100	84.94
Friedman_mse	10	84.14
	50	85.21
	100	85.01
Poisson	10	83.66
	50	85.02
	100	85.41

Discussion

- Four different ML models i..e., Multilinear Regression model, Support Vector Machine Regression model, Decision Tree Regression model, and Random Forest Regression model were tested, and the results are provided above.
- The chosen models were also subjected to hyper tunning of parameters.
- It is noted that the Random Forest Regression model produced output with comparatively higher accuracy i.e., > 84%.
- During the hyper tunning of parameters, it was found that Random Forest Regression model with n_estimators = 50 and Criterion = Absolute Error gave a better prediction with 86.04 % accuracy i.e., R^2 value.

Conclusion

Hence, the Random Forest Regression model with n_estimators = 50 and Criterion = Absolute Error can be selected as the final model for the study.