Regression

Problem

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.

Identify your problem statement - 3 stages of problem identification

Machine Learning - Number data
Supervised Learning - Given input and Output
Regression – Output is numerical

Basic info about the dataset:

total number of rows -1338

total number of columns -6

The Pre-Processing method

converting string to number – nominal data. one hot encoding

1. Simple Linear Regression : r-**score=0.7894**
2. Multiple Linear Regression : r-**score=0.7894**

3.Support Vector Machine:

Sl.No	Hyper parameter (c)	Linear R^2	Poly R^2	Rbf R^2	Sigmoid R^2
1	C=10	0.46	0.03	-0.03	0.03
2	C=100	0.62	0.61	0.32	0.52
3	C=500	0.76	0.82	0.66	0.44
4	C=1000	0.76	0.85	0.81	0.28
5	C=2000	0.74	0.86	0.85	-0.59
6	C=4000	0.74	0.86	0.87	
7	C=10000	0.74	0.85	0.87	

The SVM Regression best R^2 value is 0.87 using Rbf parameter(C=10000)

4.Decision Tree Method:

Sl.No	criterion	splitter	R^2
1	mse	best	0.70
2	mse	random	0.71
3	friedman_mse	best	0.69
4	friedman_mse	random	0.68
5	mae	best	0.67
6	mae	random	0.74

The Decision Forest Regression best R^2 value is 0.74 using

Criterion = mae(absolute_error), Splitter = Random parameter

Random Forest:

Sl.No	n_estimators	R^2
1	50	0.84
2	100	0.85
3	200	0.85
4	250	0.85

The Decision Forest Regression best R^2 value is 0.85 using n_estimators=100

Final model & justification for chosen

So far Analysis the all regression algorithm we got a best R^2 value is 0.87 using Support Vector Machine Algorithm..So we choose the final model is SVM.