**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

Problem Statement: To predict the insurance charges

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

There are totally 1338 rows (datapoints). It’s a good amount of input for the model. There are five independent variables (age, sex, bmi, children, smoker) and one dependent variable (Charges).

3.) Mention the pre-processing method if you’re doing any (like converting

string to number – nominal data)

The independent variables - sex, smoker are categorical data and needs to be converted to numerical.

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.

I used five machine learning algorithms to create different models and find R2 score based on different hyper tuning parameters. I have captured the prominent combinations for each model in the table below. The worst model is the Support Vector Machine (SVM) and the best model is the Random Forest.

5.) All the research values (r2\_score of the models) should be documented.

(You can make tabulation or screenshot of the results.)

|  |  |  |
| --- | --- | --- |
| **Model** | **Tuning Parameters** | **R2 score** |
| Multiple Linear Regression |  | 0.786511 |
| SVR | kernel=linear, C=1.0 | -0.14846 |
|  | kernel=poly, C=1.0 | -0.0797 |
|  | kernel=rbf, C=1.0 | -0.09799 |
| NuSVR | kernel=linear, C=1.0 | 0.095373 |
|  | kernel=poly, C=1.0 | -0.01476 |
|  | kernel=rbf, C=1.0 | -0.0496 |
| DecisionTreeRegressor | criterion="squared\_error",splitter="best",random\_state=42 | 0.690028 |
|  | criterion="friedman\_mse",splitter="random",random\_state=0 | 0.736502 |
|  | criterion="squared\_error",splitter="random",random\_state=0 | 0.731502 |
|  | criterion="absolute\_error",splitter="random",random\_state=0 | 0.756583 |
| Random Forest Regressor | criterion="squared\_error",n\_estimators=100, random\_state = 42 | 0.860731 |
|  | criterion="absolute\_error",n\_estimators=100, random\_state = 42 | 0.859297 |
|  | criterion="friedman\_mse",n\_estimators=100, random\_state = 42 | 0.860391 |
|  | criterion="poisson",n\_estimators=100, random\_state=42 | 0.859006 |

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

The final model I am choosing is the Random Forest Regressor. With the highlighted criterion above, it provided the better R2 score for me. I was not able to get a >90% score though.

Kindly create Repository in the name Regression

Assignment.

Upload all the ipynb and final document in the pdf

Uploaded the files in the following Github link