**Machine Learning – Regression Assignment**

**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.

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

3.) Mention the pre-processing method if you’re doing any (like converting string to number – nominal data)

4.) Develop a good model with r2\_score. You can use any machine learning algorithm; you can create many models. Finally, you must come up with final model.

5.) All the research values (r2\_score of the models) should be documented. (You can make tabulation or screenshot of the results.)

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

Kindly create Repository in the name Regression Assignment.

Upload all the ipynb and final document in the pdf

Communication is important (How you are representing the document.)

1. **Problem Statement Identification:**

The client’s requirement is to predict the insurance charges. Client has provided the data and it as age, sex, bmi, children, smoker, and charges column.

**Domain Selection:**

By Analysing the customer input, we could see that 3 columns (age, bmi, children) are numeric, and 2 columns (sex and smoker) are having nominal data and the output is numeric. Hence, we need to select the **Machine Learning Domain.**

1. **Basic information about dataset:**

User datasheet has 6 columns. It has 5 inputs and 1 output. Age, sex, bmi, children, smoker are input columns and Charges in the output column.

**Learning Selection:**

In this problem statement the requirement is clear, and the dataset is having both input and output. Hence it should undergo **supervised learning**.

**Algorithm Selection:**

Since the output is numeric continuous, we should use the **regression algorithm** under supervised learning.

1. **AI has 2 phases:**

AI model creation contains 2 phases:

1. Model Creation
2. Deployment
3. **Model Creation phases:**

Model creation has the below mentioned steps:

* + 1. Data Collection
    2. Data Preprocessing
    3. Input/Output Split
    4. Splitting Test and Train data
    5. From the Training data create the model
    6. Using the test data set predict the model
    7. Evaluate the model using Evaluation Metrics
    8. Save the best model.

1. **Deployment Phase:**

Deployment phase has the below mentioned steps:

* + 1. Load the saved model
    2. Get the inputs
    3. Predict
    4. Call to Action

1. **Algorithms for Regression:**

Since our data uses regression algorithm there are following regression algorithms it has the following algorithms

1. Linear Regression algorithm
2. Multiple Linear Regression algorithm
3. Polynomial Regression algorithm
4. **Algorithms for Regression & Classification:**
5. Support vector machine
6. Decision Tree
7. Random Forest.
8. **Categorical Data:**

In our data the 2 fields Sex and Smoker as categorical data these data are **Nominal Data** hence, we need to data preprocessing for before preprocessing. Hence we need to have a **One Hot Encoding** algorithm.

1. **Linear Regression Algorithm:**

Simple Linear Regression algorithm can be applied only when there is only one input but in our case, we are having 5 inputs hence it cannot be used.

1. **Multiple Linear Regression Algorithm:**

**Weight:**

array([[ 257.8006705 , 321.06004271, 469.58113407, -41.74825718, 23418.6671912 ]])

**Bias:**

array([-12057.244846])

**R2 value:** 0.7894790349867009

1. **Support Vector Machine:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No** | **Hyper Parameters** | **Linear** | **Poly** | **Rbf** | **Sigmoid** |
| 1 | C=1.0 | -0.01010267 | -0.07569966 | -0.08338 | -0.07543 |
| 2 | C=10.0 | 0.462468414 | 0.038716223 | -0.03227 | 0.039307 |
| 3 | C=100.0 | 0.628879286 | 0.617956962 | 0.320032 | 0.52761 |
| 4 | C=500.0 | 0.763105798 | 0.826368354 | 0.664298 | 0.444606 |
| 5 | C=1000.0 | 0.764931174 | 0.856648768 | 0.810206 | 0.287471 |
| 6 | C=2000.0 | 0.744041831 | 0.860557928 | 0.854777 | -0.59395 |
| 7 | C=3000.0 | 0.74142366 | 0.859893008 | 0.866339 | -2.12442 |

The **SVM Regression** use R2 value Poly , and hyper parameter C2000=0.86055

1. **Decision Tree:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No** | **CRITERION** | **MAX\_FEATURES** | **SPLITTER** | **R VALUE** |
| 1 | *squared\_error* | None | Best | 0.69466 |
| 2 | *squared\_error* | None | random | 0.69711 |
| 3 | *Squared\_error* | sqrt | Best | 0.6851005 |
| 4 | *Squared\_error* | Sqrt | Random | 0.69335 |
| 5 | *Squared\_error* | Log2 | Best | 0.70793 |
| 6 | *Squared\_error* | Log2 | Random | 0.69879 |
| 7 | *friedman\_mse* | None | Best | 0.68305 |
| 8 | *friedman\_mse* | None | Random | 0.71174 |
| 9 | *friedman\_mse* | sqrt | Best | 0.66711 |
| 10 | *friedman\_mse* | sqrt | Random | 0.66725 |
| 11 | *friedman\_mse* | Log2 | Best | 0.73162 |
| 12 | *friedman\_mse* | Log2 | Random | 0.62146 |
| 13 | *absolute\_error* | None | Best | 0.6680105 |
| 14 | *absolute\_error* | None | Random | 0.76551 |
| 15 | *absolute\_error* | sqrt | Best | 0.74973 |
| 16 | *absolute\_error* | sqrt | Random | 0.57414 |
| 17 | *absolute\_error* | Log2 | Best | 0.72417 |
| 18 | *absolute\_error* | Log2 | Random | 0.68936 |
| 19 | *poisson* | None | Best | 0.73342 |
| 20 | *poisson* | None | Random | 0.73735 |
| 21 | *poisson* | sqrt | Best | 0.69777 |
| 22 | *poisson* | sqrt | Random | 0.63266 |
| 23 | *poisson* | Log2 | Best | 0.72033 |
| 24 | *poisson* | Log2 | Random | 0.62721 |

Using Decision Tree algorithm R2 value 0.76551 with Criterion absolute\_error and Max\_features is none and split = random.

1. **Random Forest:**

|  |  |  |
| --- | --- | --- |
| **S.No** | ***n\_estimators*** | **R2 Value** |
| 1 | 50 | 0.849832932 |
| 2 | 100 | 0.853830791 |
| 3 | 500 | 0.853129707 |

Using Random Forest algorithm when the n\_estimators value is 100 we are getting R2 value as 0.85

**Conclusion:**

By executing different algorithms we are getting maximum R2 value of 0.860557928 only for Support Vector Machine Algorithm with the hyper parameters C=2000 hence we need to use SVM Algorithm in this case