# REGRESSION ASSIGNMENT-ML

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.

Prediction: Insurance charges

- Domain Selection: Machine Learning
- Learning Selection:Supervised Learning (Requirement is clear)
- Regression(label is in numerical values)

- 1338 rows × 6 columns in the given dataset
- I have done preprocessing method in the columns sex,smokers(converting nominal data to numerical data using (one hot encoding function )

## Multiple Linear Algorithm:

Using Multiple linear ,I got r2 value=0.7894

## Support Vector Machine Algorithm:

Using svm algorithm's hyper tuning parameters:
{linear,rbf(nonlinear),poly,sigmoid}

| Hyper parameter | Linear R2 value | Rbf R2 value | Poly R2 value | Sigmoid R2 value |
|---|---|---|---|---|
| C=10 | 0.462 | -0.032 | 0.038 | 0.039 |
| C=100 | 0.628 | 0.320 | 0.617 | 0.527 |
| C=500 | 0.763 | 0.664 | 0.826 | 0.444 |
| C=1000 | 0.764 | 0.810 | 0.856 | 0.287 |
| C=1500 | 0.764 | 0.842 | 0.858 | -0.067 |
| C=2000 | 0.744 | 0.854 | 0.860 | -0.593 |
| C=3000 | 0.741 | 0.866 | 0.859 | -2.124 |

In this algorithm,r2 value is very low

## DECISION TREE ALGORITHM:

Hyper tuning parameter:

- criterion={squared_error,absolute_error,poisson,friedman_mse}
- splitter={random,best}
- max_features={none,log2,sqrt}

| S.No | Criterion | Splitter | Max features | R2 value |
|---|---|---|---|---|
| 1. | Squared_error | Best | None | 0.682 |
| 2. | Squared_error | Random | None | 0.648 |
| 3. | Squared_error | Best | Sqrt | 0.680 |
| 4. | Squared_error | Best | Log2 | 0.688 |
| 5. | Squared_error | Random | Sqrt | 0.700 |
| 6. | Squared_error | Random | Log2 | 0.693 |
| 7. | Friedman_mse | Best | None | 0.697 |
| 8. | Friedman_mse | Random | None | 0.706 |
| 9. | Friedman_mse | Best | Log2 | 0.754 |
| 10. | Friedman_mse | Random | Log2 | 0.667 |
| 11. | Friedman_mse | Best | Sqrt | 0.785 |
| 12. | Friedman_mse | Random | Sqrt | 0.579 |
| 13. | Absolute_error | Best | None | 0.684 |
| 14. | Absolute_error | Random | None | 0.701 |
| 15. | Absolute_error | Best | Sqrt | 0.764 |
| 16. | Absolute_error | Random | Sqrt | 0.687 |
| 17. | Absolute_error | Best | Log2 | 0.686 |
| 18. | Absolute_error | Random | Log2 | 0.734 |
| 19. | Poisson | Best | None | 0.717 |
| 20. | Poisson | Best | Log2 | 0.746 |
| 21. | Poisson | Best | Sqrt | 0.697 |

|  |  |  |  |  |
|---|---|---|---|---|
| 22. | Poisson | Random | None | 0.674 |
| 23. | Poisson | Random | Log2 | 0.725 |
| 24. | Poisson | Random | Sqrt | 0.751 |

In this algorithm, we didn’t get a good model. (r2 value is low)

## RANDOM FOREST ALGORITHM

Hyper tuning parameters:

- criterion={squared_error,absolute_error,poisson,friedman_mse}
- n_estimators={50,100}
- max_features={none,log2,sqrt}

| S.No | Criterion | n_estimators | max_features | r2 value |
|---|---|---|---|---|
| 1. | Squared_error | 50 | None | 0.852 |
| 2 | Squared_error | 100 | Sqrt | 0.864 |
| 3 | Squared_error | 100 | None | 0.856 |
| 4 | Squared_error | 50 | Sqrt | 0.867 |
| 5 | Squared_error | 50 | Log2 | 0.866 |
| 6 | Squared_error | 100 | Log2 | 0.870 |
| 7 | Absolute_error | 50 | None | 0.849 |
| 8 | Absolute_error | 50 | Sqrt | 0.872 |
| 9 | Absolute_error | 50 | Log2 | 0.865 |
| 10 | Absolute_error | 100 | None | 0.853 |
| 11 | Absolute_error | 100 | Sqrt | 0.869 |
| 12 | Absolute_error | 100 | Log2 | 0.875 |
| 13 | mse | 50 | None | 0.849 |
| 14 | Mse | 50 | Sqrt | 0.869 |

|  |  |  |  |  |
|---|---|---|---|---|
| 15 | Mse | 50 | Log2 | 0.868 |
| 16 | Mse | 100 | None | 0.853 |
| 17 | Mse | 100 | Sqrt | 0.869 |
| 18 | Mse | 100 | Log2 | 0.870 |
| 19 | poisson | 50 | None | 0.855 |
| 20 | Poisson | 50 | Sqrt | 0.864 |
| 21 | poisson | 50 | Log2 | 0.869 |
| 22 | Poisson | 100 | None | 0.857 |
| 23 | Poisson | 100 | Sqrt | 0.869 |
| 24 | poisson | 100 | Log2 | 0.870 |

- In all these above algorithms , r2 value is low compared to random forest algorithm. So I didn’t proceed to the next phase.

- Using random forest algorithm’s parameter (criterion=absolute_error,n_estimators=100,max_features=log2). I got r2=0.875 (nearly to 1). This is a better model.so I proceed to the phase II (i.e,Deployment phase)