**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

**Dataset:**

Git Hub Link: <https://raw.githubusercontent.com/RamishaRaniK/dataset/main/insurance_pre.csv>

**Identified Problem Statement:**

To build a Boosting Algorithm - regression model that predicts the insurance charges incurred by individuals based on their health related parameters

**Basic Info about the Dataset:**

 Total number of rows: 1338

 Total number of columns: 6

**Pre-processing Methods Used**

Converted categorical 'smoker' and ‘Sex’ column to one-hot encoding

**Model Development & Evaluation (R² Score)**

1. **Ada Boost**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.NO** | **estimator** | **n\_estimators** | **learning\_rate** | **loss** | **r\_score** |
| 1 | None | 50 | 1.0 | Linear | 0.9943019088865389 |
| 2 | None | 50 | 1.0 | square | 0.9959198002018387 |
| 3 | None | 50 | 1.0 | exponential | 0.9950350501619002 |

The best model is **Ada Boost with loss=’square’**, achieving the highest R² score of **0.9959.**

1. **XG Boost**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S.NO** | **n\_estimators** | **max\_depth** | **eta** | **subsample** | **colsample\_bytree** | **r\_score** |
| 1 | 500 | 5 | 0.05 | 0.9 | 0.6 | 0.9961899518966675 |
| 2 | 800 | 6 | 0.2 | 0.8 | 0.7 | 0.9974551200866699 |
| 3 | 1200 | 4 | 0.15 | 0.75 | 0.9 | 0.9992831945419312 |
| 4 | 600 | 3 | 0.1 | 0.85 | 0.5 | 0.9978577494621277 |
| 5 | 1000 | 8 | 0.07 | 0.65 | 0.8 | 0.9954016804695129 |

The best model is **XG Boost** with n\_estimators=1200, max\_depth=4, eta=0.15, subsample=0.75, colsample\_bytree=0.9, achieving the highest R² value of **0.9992831945419312**

1. **LG Boost**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **S.NO** | **boosting\_type** | **n\_estimators** | **learning\_rate** | **num\_leaves** | **subsample** | **max\_depth** | **r\_score** |
| 1 | gbdt | 500 | 0.05 | 31 | 0.8 | 6 | 0.9961899518966675 |
| 2 | gbdt | 300 | 0.05 | 31 | 0.8 | 6 | 0.9985738833965891 |
| 3 | gbdt | 100 | 0.05 | 31 | 0.8 | 6 | 0.9988740719992442 |
| 4 | dart | 500 | 0.05 | 31 | 0.8 | 6 | 0.9981035923851933 |
| 5 | dart | 300 | 0.05 | 31 | 0.8 | 6 | 0.9922518321794354 |
| 6 | dart | 100 | 0.05 | 31 | 0.8 | 6 | 0.9182300985961904 |
| 7 | goss | 500 | 0.05 | 31 | 0.8 | 6 | 0.9985327283679531 |
| 8 | goss | 300 | 0.05 | 31 | 0.8 | 6 | 0.9987631160180305 |
| 9 | goss | 100 | 0.05 | 31 | 0.8 | 6 | 0.9990232589614273 |

The best model is **LG Boost Regressor** with parameters boosting\_type='goss', n\_estimators=500, learning\_rate=0.05, num\_leaves=31, max\_depth=6, subsample=0.8 achieving the highest R² value of **0.9990232589614273.**

**The Final Machine Learning Best Method of Boosting Algorithm Regression**

**XG Boosting Regression is selected as the best model** for predicting insurance charges due to its perfect performance (R² = **0.9992831945419312**)