Name : Nayanathara P.M.C.

Index No : 210417X

CS3111 - Introduction to Machine Learning

Lab 02 – Regression

➢ *Introduction*

The challenge of this lab was to analyze the **Sustainable Urban Living** Kaggle competition dataset. The task was to develop a machine learning model that predicts the 'Habitability_score' for each property.

➢ *Data Set*

- Two separate data sets to train and test.
- Total number of features : 15
- Number of train records : 31599
- Number of test records : 7900

➢ *Data Analysis & Preprocessing*

After loading and analyzing the dataset, the following data preprocessing techniques were performed on the dataset.

- Visualizing the correlation between each pair of features.
- Checking the percentage of missing values of each feature.
- Categorizing numerical and categorical features.
- Visualizing the distribution of outliers.
- Imputing the numerical missing values with the mean of each feature.
- Standardizing the numerical features.
- Imputing the missing values in categorical features with the mode.
- Dropping ''Id'' column.
- Applying One Hot Encoding, Label Encoding appropriately to convert the categorical features to numerical values.
- Separating X_train, X_test and y_train data.

➢ *Model Selection*

Approach 1 :

- Firstly, five machine learning models namely, **Linear Regression, Lasso, ElasticNet, DecisionTree, Random Forest Regressor and Gradient Boosting Regressor** and a grid containing the hyperparameters for each model was defined.
- Each model is tunned using grid-search method by five-fold cross validation taking negative mean squared error as the scoring method. By that way, the best set of estimators for each model is found.
- Thereafter, these best models are again evaluated to find the best performing model for the regression problem. It resulted in the **Random Forest Regressor** being the best performing model.
- This model was used to predict the target values for the test dataset.

Approach 2 :

- A neural network architecture was defined having an output layer with a single neuron for regression. Then the model is compiled and evaluated using the mean squared error as the loss function.
- The train data is split into train and validation sets and trained the model and evaluated using the train and validation losses.

➢ *Evaluation Metrics*

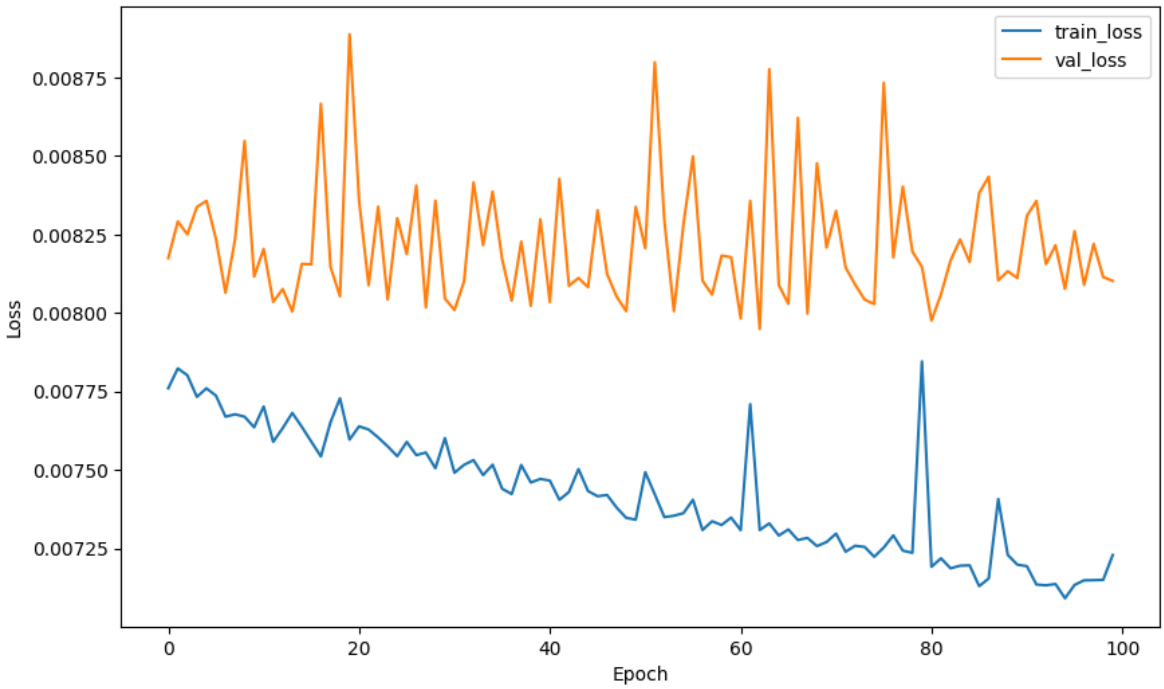
The trained models were evaluated using Negative Mean Squared Error and R2-score metrics to assess the accuracy of each model.

1. **Negative Mean Squared Error** : Negated value of MSE (Mean of the squared values between the actual and predicted values)
2. **R2-score** : This measure is used to evaluate the performance of regression models. It provides insights into how well the regression model fits the observed data.

Comparison of Results

	Linear Regression	Lasso Regression	ElasticNet Model	DecisionTree Model	Random Forest Regressor	Gradient Boosting Model
NMSE	-84.44	-84.44	-84.44	-43.85	-35.47	-45.14
R2-score	0.57	0.57	0.57	0.78	0.82	0.77

The accuracy of the neural network model is measured using the '**mean squared error**' and '**mean squared logarithmic error**'.



➢ *Conclusion*

Based on the evaluation metrics, the RandomForestModel demonstrates a better performance compared to the other models for predicting the 'Habitability_Score'. The RandomForestModel achieves a lower value for MSE and a high value for R2-score, indicating better accuracy and closer predictions to the actual values.

➢ *Final Score & Ranking of the Final Submission*