CS3111- Introduction to Machine Learning

Lab 02-Regression

Lab report

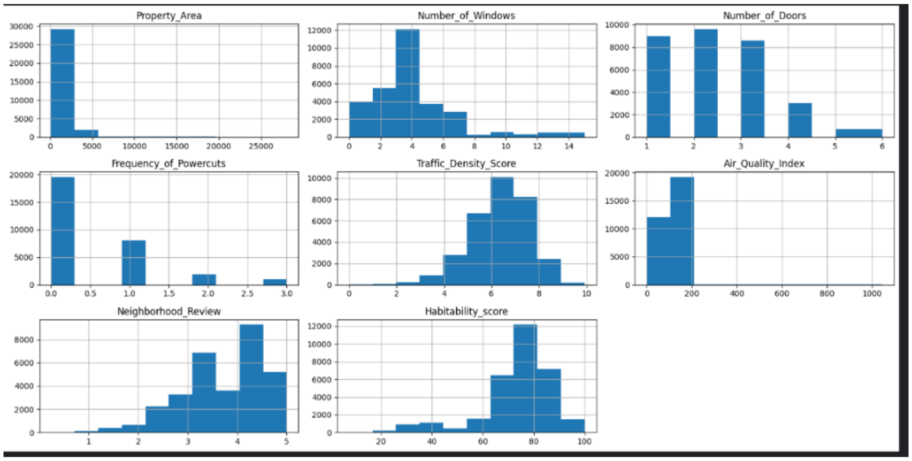
Index : 210202J Name : A.H.H.Hana

Introduction

In this lab, we are given a dataset to assess a property's true "habitability" based solely on descriptions. This is a Kaggle competition and evaluation is mainly based on Root Mean Squared Error. https://www.kaggle.com/competitions/ml-olympiad-sustainable-urban-living/overview

Visualization

Data can be visualized by a histogram as below:



Models Building and Evaluation

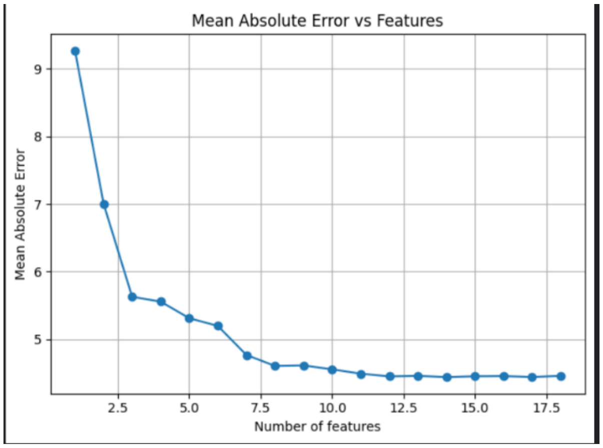
Following Regression methods were used to build models. Models are trained with data frame given as train data. Train data is split into 2 where one set is used to train and the other to test. Each models's performance is measured by calculating the Root Mean Squared Error, Mean Squared Error, Mean Absolute error and R2 scores.

And the model with least RMSE value is selected to predict the habitability scores of the given data set.

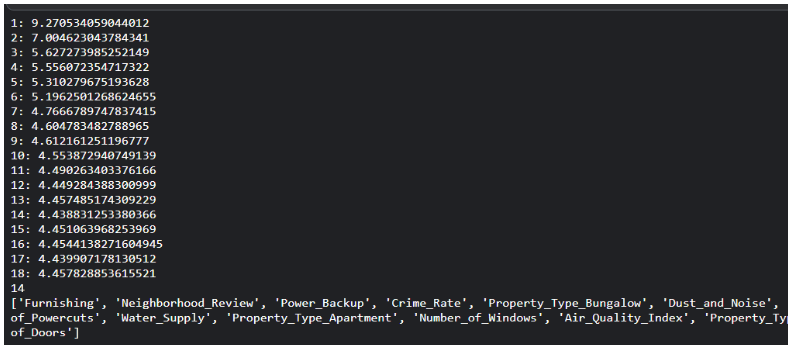
1. Linear Regression
2. Random Forest
3. Gradient Boosting
4. Support Vector Regression
5. ElasticNet Regression
6. k-Nearest Neighbors Regression
7. Decision Tree Regression
8. Hyperparameter tuned XGB Regression

Then, by performing feature selection using a forward selection method combined with Random Forest regression, the performance of the models are evaluated using Mean Absolute Error (MAE).

The Mean Absolute Error vs Number of Features is plotted as follows:

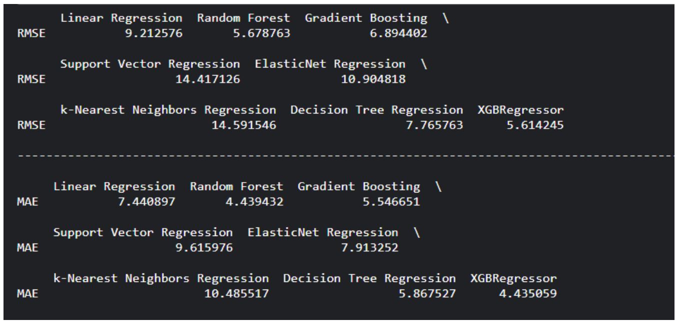


Evaluation of each model using forward selection method is shown below:

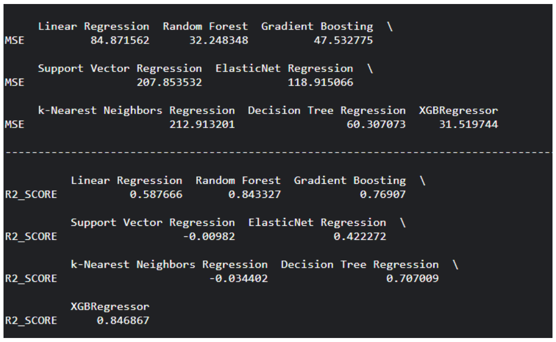


```
1: 9.270534059044012
2: 7.004623043784341
3: 5.627273985252149
4: 5.556072354717322
5: 5.310279675193628
6: 5.1962501268624655
7: 4.7666789747837415
8: 4.604783482788965
9: 4.612161251196777
10: 4.553872940749139
11: 4.490263403376166
12: 4.449284388300999
13: 4.457485174309229
14: 4.438831253380366
15: 4.451063968253969
16: 4.4544138271604945
17: 4.439907178130512
18: 4.457828853615521
14
['Furnishing', 'Neighborhood_Review', 'Power_Backup', 'Crime_Rate', 'Property_Type_Bungalow', 'Dust_and_Noise',
of_Powercuts', 'Water_Supply', 'Property_Type_Apartment', 'Number_of_Windows', 'Air_Quality_Index', 'Property_Ty
of_Doors']
```

After building models, the performance of each model is calculated using multiple evaluation metrics



```
      Linear Regression  Random Forest  Gradient Boosting  \
RMSE           9.212576       5.678763           6.894402   

      Support Vector Regression  ElasticNet Regression  \
RMSE                  14.417126              10.904818   

      k-Nearest Neighbors Regression  Decision Tree Regression  XGBRegressor  
RMSE                       14.591546                  7.765763      5.614245  

---------------------------------------------------------------------------------------------

     Linear Regression  Random Forest  Gradient Boosting  \
MAE           7.440897       4.439432           5.546651   

     Support Vector Regression  ElasticNet Regression  \
MAE                   9.615976               7.913252   

     k-Nearest Neighbors Regression  Decision Tree Regression  XGBRegressor  
MAE                       10.485517                  5.867527      4.435059  
```



```
     Linear Regression  Random Forest  Gradient Boosting  \
MSE          84.871562      32.248348          47.532775   

     Support Vector Regression  ElasticNet Regression  \
MSE                 207.853532             118.915066   

     k-Nearest Neighbors Regression  Decision Tree Regression  XGBRegressor  
MSE                      212.913201                 60.307073     31.519744  

-------------------------------------------------------------------------------------

          Linear Regression  Random Forest  Gradient Boosting  \
R2_SCORE           0.587666       0.843327            0.76907   

          Support Vector Regression  ElasticNet Regression  \
R2_SCORE                   -0.00982               0.422272   

          k-Nearest Neighbors Regression  Decision Tree Regression  \
R2_SCORE                       -0.034402                  0.707009   

          XGBRegressor  
R2_SCORE      0.846867  
```

XGBoost has a lowest RMSE. MSE , MAE and highest R2 Score for both test and train data. Selecting the model with the lowest RMSE, MSE, and MAE, and the highest R2 score ensures that the model provides accurate predictions with good explanatory power, minimizing errors and maximizing the proportion of variance explained.

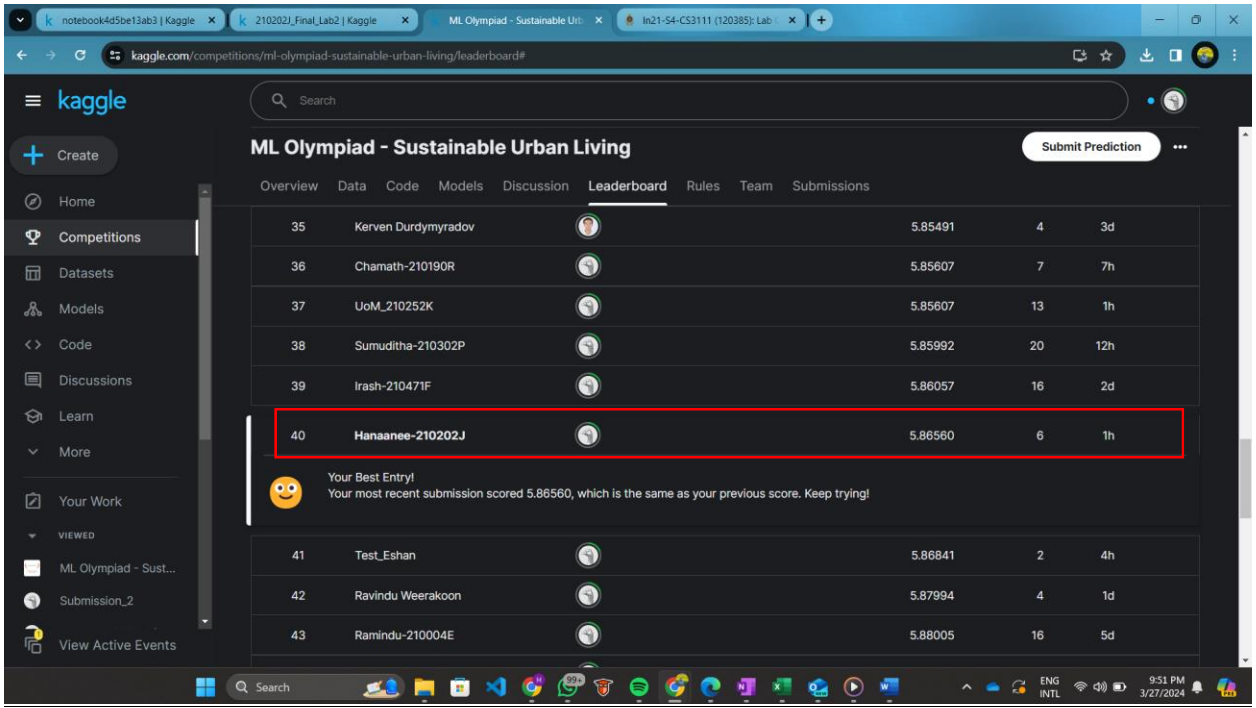
Finally, the best model, **XGB Regressor** with least Root Mean Squared Error is selected to predict the Habitability scores.

Conclusion

Based on the results, it can be observed that the XGB Regressor achieved lower RMSE,MAE,MSE along with highest R2 score compared to all the other Regressor models. It can be concluded that the XGB Regressor model outperformed all the other models for this specific dataset and task.

XGBoost is an ensemble method that combines multiple weak learners (decision trees) to create a strong learner. This often leads to better performance compared to individual models, especially when dealing with complex relationships in the data. XGBoost provides numerous hyperparameters that can be tuned to optimize model performance. With effective hyperparameter tuning strategies, XGBoost can often achieve better results compared to models with fewer hyperparameters or less flexibility. Since we have used RandomizedSearch Hyperparameter tuning method, it is possible for the XGB Regressor model to outperform Random Forest and the other models trained in this case.

Leaderboard with Score



ML Olympiad - Sustainable Urban Living

35	Kerven Durdymyradov	5.85491	4	3d
36	Chamath-210190R	5.85607	7	7h
37	UoM_210252K	5.85607	13	1h
38	Sumuditha-210302P	5.85992	20	12h
39	Irash-210471F	5.86057	16	2d
40	**Hanaanee-210202J**	5.86560	6	1h
41	Test_Eshan	5.86841	2	4h
42	Ravindu Weerakoon	5.87994	4	1d
43	Ramindu-210004E	5.88005	16	5d

Your Best Entry!
Your most recent submission scored 5.86560, which is the same as your previous score. Keep trying!