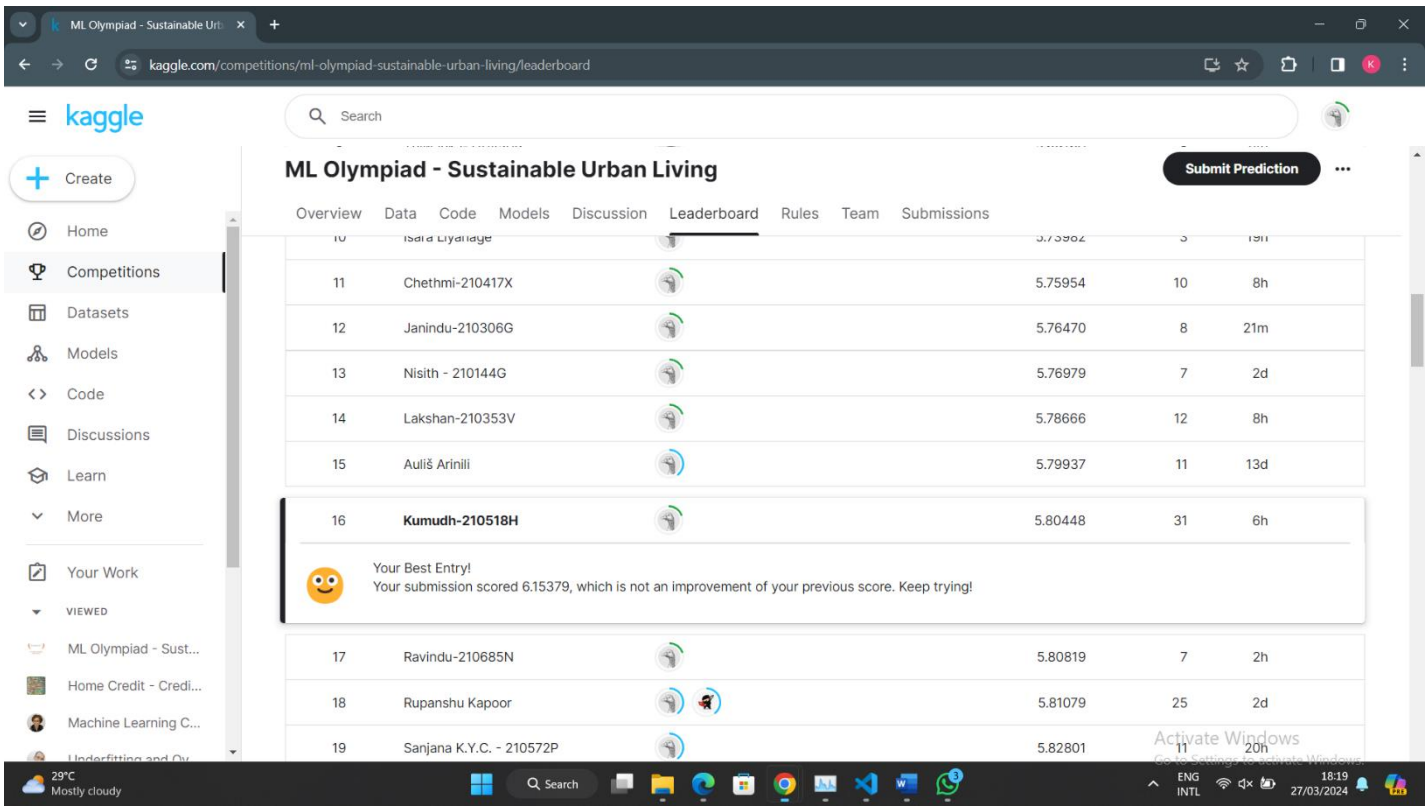
In20-S4-CS3111 – Introduction to Machine Learning

Lab 02 - Regression

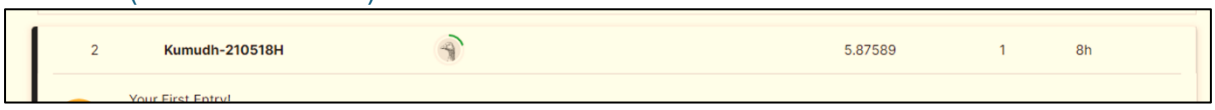
NAME : K.S. RANASINGHE **INDEX NO. : 210518H**

Final Rank/Score

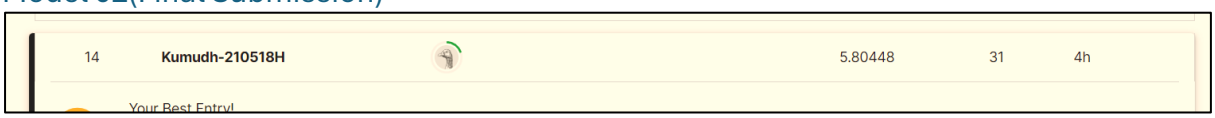


For this report I will do a comparison between the first submission I made, and the final submission made using various evaluation metrics to understand how these two models perform against each other. Given below are the scores obtained by the two models for the submissions made to the Kaggle Competition.

Model 01(First Submission)



Model 02(Final Submission)



As you can see from above the score improves from 5.87589 to 5.80448 when we go from model 01 to 02. This is on the unseen test.csv. We will be evaluating on the train.csv dataset. First, we will see how these models work. The differences are highlighted in yellow for the second model.

Model 01

- Missing values were handled using a prediction model rather than imputation. Used Lienear Regressor as the model for predicting missing values.
- Encoding done using Ordinal encoder.
- No Feature Scaling.
- No Feature Selection.
- No Dimensionality Reduction.
- The ML model was made using ensemble methods. Created a Stacked Model using XGBRegressor, RandomForestRegressor and ExtraTreesRegressor as base models and Linear Regressor as the meta model.
- No Hyper parameter tuning.

Model 02

- Missing values were handled using a prediction model rather than imputation. Used Random Forest Classifier as the model for predicting missing values.
- Encoding done using Ordinal encoder.
- Feature Scaling was done.
- No Feature Selection.
- No Dimensionality Reduction.
- The ML model was made using ensemble methods. Created a Blended Model using XGBRegressor, RandomForestRegressor and ExtraTreesRegressor as base models and Lasso Regressor as the meta model.
- Hyper parameter tuning was done for the base models.

Now that we have an understanding of the two models, we will delve onto the evaluation of the performance of the two models on the train.csv dataset.

	Model 01	Model 02
Root Mean Square Error	5.497825695324197	5.386956791860444
Mean Absolute Error	4.292745969987085	4.143820939580833
Mean Square Error	30.22608737616699	29.01930347737137
R-Squared	0.8410428074185596	0.8510758765918485
Mean Absolute Percentage Error	0.0641511156662835	0.06144337899991974
Explained Variance Score	0.8410428074185596	0.8512308082536888
Max Error	37.26656870944318	40.89042940836921
Mean Squared Log Error	0.00742650810532961	0.0073521119987292665

Root Mean Square Error (RMSE)

Model 01: 5.497826 Model 02: 5.386957

RMSE is calculated by squaring the errors and then taking the root to compensate for negative value errors. Lower values indicate better performance. Model 02 has a slightly lower RMSE, suggesting it is better at predicting the target variable.

Mean Absolute Error (MAE)

Model 01: 4. 292746 Model 02: 4. 143821
MAE represents the average magnitude of the errors in predictions. Again absolute values taken to compensate for negative errors. Like RMSE, lower values are desirable. Model 02 again shows a lower MAE.

Mean Squared Error (MSE)

Model 01: 30. 226087 Model 02: 29. 019303
MSE measures the average of the squares of the errors. Lower values indicate better performance, and Model 02 has a lower MSE.

R-Squared (R2)

Model 01: 0. 841043 Model 02: 0. 851076
R-squared represents the proportion of the variance in the dependent variable that is predictable from the independent variables. Values closer to 1 indicate a better fit. Model 02 has a higher R2, suggesting it explains more variance in the data.

Mean Absolute Percentage Error (MAPE)

Model 01: 0. 064151 Model 02: 0. 061443
MAPE expresses errors as a percentage of the actual values. Lower values indicate better accuracy. Model 02 shows a slightly lower MAPE.

Explained Variance Score

Model 01: 0. 841043 Model 02: 0. 851231
EVS indicates the proportion of the variance in the dependent variable that is explained by the independent variables. Model 02 has a higher EVS which is desirable.

Max Error

Model 01: 37. 266568 Model 02: 40. 890429
Max Error represents the maximum residual error. Lower values are preferred. Model 01 has a lower max error. Unhandled Outliers contribute to this error.

Mean Squared Log Error (MSLE)

Model 01: 0. 007426 Model 02: 0. 007352
MSLE measures the mean of the squared differences between the natural logarithm of the predicted values and the natural logarithm of the observed values. Lower values indicate better performance, and both models have similar MSLE.

As you can see above in all the evaluation metrics Model 02 outperforms Model 01. This is understandable because Model 02 was built by improving possible issues that I identified in Model 01. Possible reasons could be the use of Classifier instead of a Regressor for predicting missing values, Scaling the processed dataset and hyper parameter tuning. Another reason could be that Model 01 is overfitting therefore performing poorly on unseen data. This issue is solved in Model 02 by the use of regularization from Lasso regressor as the meta model.