Milestone – 4

* Creating multiple models that are widely used in machine learning.
* Evaluating model based on different evaluation parameters.
* Building a conclusive visual report on evaluation parameter.
* A brief and inferences about each model based on all evaluation parameter.

**Machine Learning models build:**

Since this is a regression problem, various regression models that are widely used are as listed below:

|  |  |
| --- | --- |
| 1 | Linear Regression |
| 2 | Lasso |
| 3 | Ridge |
| 4 | Decision tree Regressor |
| 5 | KNN Regressor |
| 6 | Random Forrest Regressor |
| 7 | SVM |
| 8 | Gradient Boost Regressor |
| 9 | XG Boost Regressor |
| 10 | ADA Boost Regressor |

**Evaluation Parameters:**

In regression problems, there are several evaluation parameters commonly used to assess the performance of the models. Here are some of the key evaluation parameters for regression:

**Mean Absolute Error (MAE):**

MAE measures the average absolute difference between the predicted and actual values. It provides an understanding of the average magnitude of errors.

It is calculated as the mean of the absolute differences between predicted and actual values.

Lower values of MAE indicate better model performance.

**Mean Squared Error (MSE):**

MSE measures the average squared difference between the predicted and actual values. It penalizes larger errors more than MAE.

It is calculated as the mean of the squared differences between predicted and actual values.

Lower values of MSE indicate better model performance.

**Root Mean Squared Error (RMSE):**

RMSE is the square root of MSE and provides an easily interpretable measure of the average error.

It is calculated as the square root of MSE.

Lower values of RMSE indicate better model performance.

**R-squared (R²) or Coefficient of Determination:**

R-squared measures the proportion of the variance in the dependent variable that can be explained by the independent variables.

It is calculated as 1 minus the ratio of the sum of squared errors to the total sum of squares.

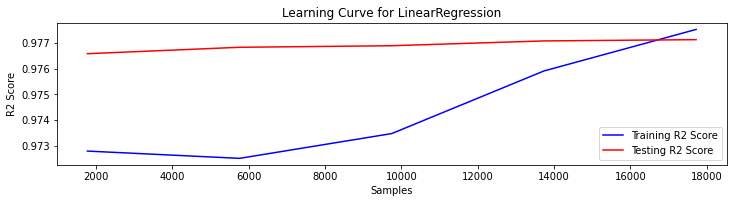
Higher values of R-squared indicate better model performance, with a maximum value of 1 indicating a perfect fit.

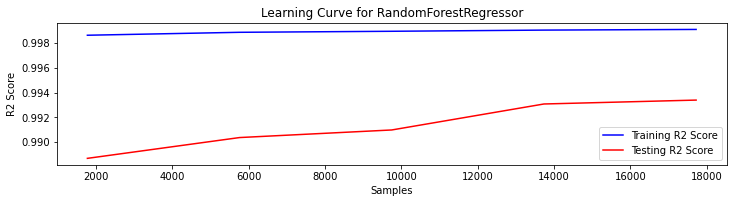
**Results:**

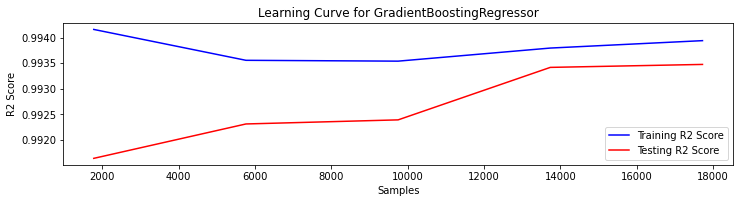
|  |  |  |
| --- | --- | --- |
| **ML Model** | **R2\_score** | **RMSE** |
| Gboost | 0.994012 | 901.5113 |
| Xgboost | 0.993911 | 929.904 |
| Random Forrest | 0.993629 | 929.904 |
| Decision tree | 0.988906 | 1227.041 |
| Lasso | 0.977939 | 1730.365 |
| Ridge | 0.977936 | 1730.365 |
| Linear regression | 0.977936 | 1730.475 |
| Adaboost | 0.976961 | 1768.284 |
| KNN | 0.638951 | 7000.122 |
| SVM | -0.00011 | 11650.51 |

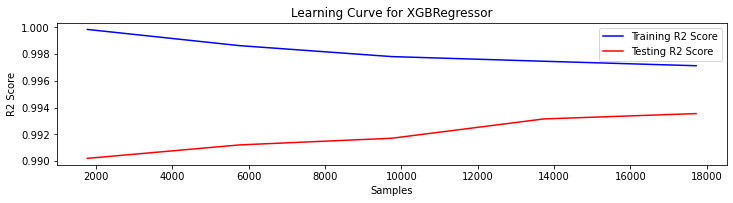
* Based on the above results these are the four ML algorithms that perform the best,
  + Gradient Boost Regressor
  + Random Forrest Regressor
  + XGBoost Regressor
  + Linear Regression

**Learning Curve Analysis:**









**Feature Importance Plot:**

To plot feature importance’s in XGBoost, one can use the plot\_importance function from the xgboost library. This function provides a bar plot showing the relative importance of each feature in the trained model.

**Importance\_type:**

This parameter defines the method of plotting the importance of each feature.

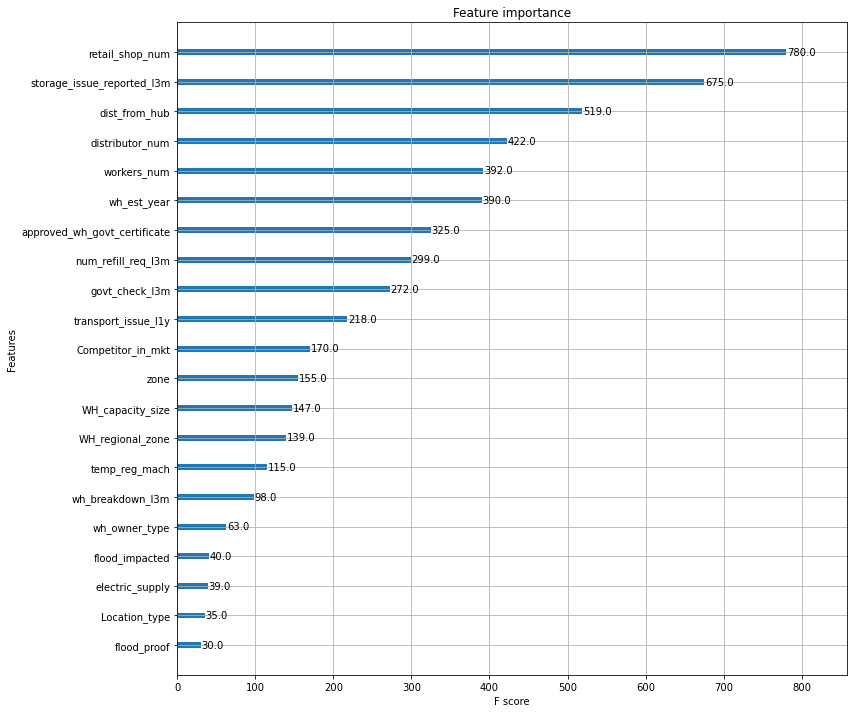
**Gain:**

“The Gain implies the relative contribution of the corresponding feature to the model calculated by taking each feature’s contribution for each tree in the model. A higher value of this metric when compared to another feature implies it is more important for generating a prediction.



**Weight:**

The Frequency (R)/Weight is the percentage representing the relative number of times a particular feature occurs in the trees of the model.



* From the importance plot we can see that following features are having high predicting power:
  + **storage\_issue\_reported\_l3m**
  + **wh\_est\_year**
  + **approved\_wh\_govt\_certificate**
  + **retail\_shop\_num**
  + **temp\_reg\_mach**
  + **workers\_num**
  + **num\_refill\_req\_l3m**

**Conclusion:**

* Tree based Machine learning models are performing very well, with RMSE as low as 901.
* Learning curve analysis shows that model is not overfitting.
* All though accuracy is very high, the feature importance plot shows that only few features are contributing to the accuracy.

THANK YOU