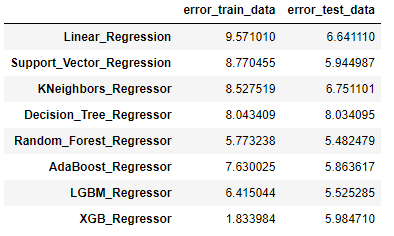
Errors of various models



* Random Forest Regressor had least test error when compared to others. LGBM Regressor had test error close to Random Forest Regressor and is many times faster. **When dataset is huge I recommend LGBM\_Regressor, else go with Random\_Forest\_Regressor**.

Overview of ML models

Linear Regression

* Assumptions

1. Linear relationship: There exists a linear relationship between the independent variable and dependent variable
2. **Homoscedasticity: Variance of residuals is constant along the values of dependent variable**
3. No Multicollinearity: The independent variables are not correlated
4. Normal Residuals: Residuals follow normal distribution
5. No autocorrelation: Observations of the error terms are uncorrelated with each other

* Advantages

1. Performs exceptionally well when there is linear relationship between independent variables and dependent variable
2. Simple to implement and easier to interpret

* Drawbacks

1. Perform poorly when assumptions are not met
2. Not capable to capture more complex patterns
3. Sensitive to outliers

Support Vector Regression

* Kernels take data as input and transform them. A hyperplane with maximum margin is selected for predictions
* Has no assumptions on Datasets
* More effective in high dimensional spaces
* Robust to outliers
* Not suitable for large datasets
* Difficult to understand and interpret

KNN Regression

* Average of targets of K nearest neighbors is taken as prediction for new data point
* Assumes data points which exist in close proximity to each other are similar
* Immediately adapts to changes as we collect new training data
* Sensitive to outliers and irrelevant features
* Not suitable for high dimensional data

Decision Tree Regressor

* A tree like structure containing root node, internal nodes, branches, leaf nodes is made. Each non-leaf node represents a condition on a feature. Prediction depends on leaf node corresponding to new data point
* Requires less effort for data preparation during pre-processing
* One of the quickest ways to identify relationships between variables and the most significant variable
* Decision trees are prone to overfitting
* A small change in data tends to cause a big difference in the tree structure

Random Forest Regressor

* Multiple decision tree are made using bootstrapping, random feature selection. Average of predictions from each tree is taken for final prediction.
* Requires less effort for data preparation during pre-processing
* Reduces overfitting as it creates many trees on the subset of the data and combines the output of all the trees
* Robust to outliers
* Is stable. A small change in data doesn’t affect predictions of algorithm
* Requires more computational power and time to train

AdaBoost Regressor

* Builds decision stumps sequentially. Each stumps corrects errors of previous stump. The amount of say each stump have on final prediction depends on total error of the stump.
* Less prone to overfitting
* Sensitive to outliers and noise
* Slower compared to other boosting algorithm

LGBM Regressor, XGB Regressor

* Builds decision trees sequentially. Each tree predicts error of previous estimator. Final prediction is sum of 1st estimator and all the errors predicted multiplied with learning rate
* Both are fast but LGBM Regressor is faster than XGB Regressor as LGBM uses leaf-wise growth and XGB use depth-wise growth
* Better performance comes at cost of overfitting