Advance House Price Prediction Report

**Data Source: Kaggle**

**The dataset has record of 1460 houses which has 81 unique feature**

**Target: Based on these features predict the price**

# Data Preprocessing

**Challenges:**

1. **Numerical Features: The dataset has 38 numerical feature**
2. Missing values:

* Problem: There are total 3 categories which has missing values. These values has relationship with target values.
* Approach: Create 3 new columns, fill 0 where value is not missing and fill 1 where value is missing. Fill the Nan values with the median.
* Status: This approach is effective.

1. Outliers:

* Problem: When compare with Sale price Most of the numerical feature has outliers
* Approach: Used IQR to find the Outliers and masked the outliers with lower limit and upper limit
* Status: This approach is not effective. (NOTE: Model is not giving the generalized prediction)

1. Distribution:

* Problem: The data is skewed
* Approach: Log Transformation, Square root Transformation, Box-cox Transformation.
* Status: Log Transformation works better than other 2. (Gives the highest accuracy )

1. **Categorical Features: The dataset has 43 categorical feature**
2. Missing Values:

* Problem: There are 11 feature which has missing values
* Approach: Fill nan values with a new Category Missing
* Status: Effective

1. Encoding:

* Problem: Categorical Feature
* Approach: one hot encoding, Target Guided encoding (will groupby the categorical feature with the mean Sale price and give each category a rank based on the mean price).
* Status: Target Guided encoding is Effective (Performed slightly better than one hot encoding)

# Feature Selection

* Problem: Dataset has 81 Feature
* Approach: Correlation with sale price, Variance Threshold, Mutual information.
* Status: Not effective (Removed dimension did not give a better result)
* Scope of improving the approach is high.

# Outliers Detection

* Problem: Most of the Numerical feature has outliers.
* Approach: Used IQR to detect the outliers and replace them with lower limit and upper limit.
* Status: Not effective (Did Perform well on training and testing, Did not increased the Rank in Competition)

# Transformation:

* Problem: Every feature is not at the same scale.
* Approach: Min Max Scaler, Standard Scaler, box-cox Transformation, Log Transformation.
* Status: Log Transformation is effective. (Gives a slightly better results)
* Scope of improving the approach is High

# Model Building

1. Linear Regression:
   1. Training Root Mean Squared Error – 0.122
   2. Test score – 0.132
   3. Score on kaggle – 0.1334
2. Support Vector Regression (kernel = polynomial, degree=4):
   1. Training Root Mean Squared Error – 0.116
   2. Test score – 0.124
   3. Score on kaggle – 0.1315
3. XGB Regressor:
   1. Training Root Mean Squared Error – 0.122
   2. Test score – 0.132
   3. Score on kaggle – 0.1438
   4. Scope to perform better – True
4. Random Forest Regressor:
   1. Training Root Mean Squared Error – 0.105
   2. Test score – 0.1470
   3. Score on kaggle – 0.1477
   4. Scope to perform better – True
5. ANN: (5 Hidden layer)
   1. Training Root Mean Squared Error – 0.099
   2. Test score – 0.015
   3. Score on kaggle – 0.136
6. Voting( Linear Regression, SVR and XGB)
   1. Training Root Mean Squared Error – 0.1083
   2. Test score – 0.1111
   3. Score on kaggle – 0.1243

## PARTICIPANTS: 5053

## RANK: 714