**House Price Prediction - Ames, Iowa**

**Domain and Context :**

As part of this capstone project we have taken the Real Estate use case of predicting the house prices. Determining the sale price of the house is very important nowadays as the price of the land and price of the house increases every year. The price of house helps the buyer to know the cost price of the house and also the right time to buy it. There are several factors that affect the price of the house such as the physical condition, location, landmark etc.

**Pre-Processing:**

Looking at the dataset which has a mix of Numeric and Categorical variables it’s important to follow some of the pre-processing steps before diving into the right algorithms and predictions. Here are some the ways to use

1. Filling NaN Values
2. Dummy Variables
3. Stemming

**Feature Engineering**

1. Reduce the number of categorical variables whenever possible / appropriate, as each categorical variable must be converted into multiple dummy variables for regular multiple linear regression models (OLS) and regularized linear regression models
2. Remove trivial variables of very low prediction value.
3. Adjust variables as required to ensure that their values or types are fit for regression purposes,

**Algorithms Used and Reasoning:**

1. **Linear Regression**:

Linear Regression was used for finding a relationship between two continuous variables, using One variable as predictor or independent, and the other variable as variable response or dependent.

1. **Ridge Regression:**

Ridge Regression was one another algorithm that was used but it suffered from multicollinearity. Because of the square estimates their variances were so large that they were far from the true value.

1. **Gradient Boosting**:

Gradient boosting was another technique used with decision tree regressors that boosted the weak predictive models. Gradient boosting leveraged the residual patterns and strengthed a model with weak predictions and made it better.

1. **Lasso Regression:**

Least Absolute Shrinkage and Selection Operator is another algorithm similar to Ridge Regression, which penalizes the absolute size of the coefficients of regression.

**RESULT**

To train the dataset and make predictions separately, LASSO (least absolute shrinkage and selection operator) and Gradient boosting regression models were used. The scores tht were achieved were

Lasso Score: 0.1115(0.0078)

GBoost Score: 0.1159(0.0088)

After this they stacked the 2 models ,trained and predicted.. here are the scores

RMSLE score on the train data: 0.06870805

Accuracy score: 0.9795672