**CS-559-House Price Prediction(Report)**

**Abstract:**

When you ask a home buyer to describe their ideal home, they are unlikely to start with the basement ceiling height or the closeness to an east-west railroad. However, the data from this playground competition shows that far more factors influence price negotiations than the number of bedrooms or the presence of a white-picket fence. However, the data in this thesis shows that the number of bedrooms and floors have a greater impact on the price of a home. I also want to use this dataset to anticipate an acceptable home price based on these attributes of the properties.

**Objective:**

Our project's purpose is to fit various machine learning models to anticipate home prices and determine which characteristics of the house have the most impact on housing prices.

**Exploratory Data Analysis:**

Scatter Plot for dependent:

Scatter plot for the dependent variable that is Sales Price where we observe that most of the points are assembled on the bottom. And there seems to be no large outliers in the sale price variable

Chart, scatter chart

Description automatically generated

Correlation:

This dataset comprises 79 explanatory variables that cover practically every facet of Ames, Iowa's residential dwellings. We can observe that the “OverallQual” seems to be directly correlated to the Sale Price with the value of 0.795437. Secondly, the continuous numeric values seem to be most affected in the correlation.

Table

Description automatically generated

Visualization:

From the plot below we observe that the Sale price of the house increases exponentially with respect to the overall quality of the house.

Chart

Description automatically generated

Whereas from the below plot the same insights can be obtained from the graph for total number of rooms above ground where similarly the Sale price have been increasing with respect to the total number of rooms above ground

Chart, bar chart

Description automatically generated

**Overview:**

We used Linear Regression, Gaussian Process, Random Forest, and XGBoost to develop this model. FireplaceQu, Alley, Id, PoolQC, Fence, and MiscFeature were omitted first, followed by obtaining correlation between two variables to verify the strongly correlated variable with its dependent variable, and finally removing all the least important columns from the dataset.

On the scatter plot points are assembled on the bottom, and outlines are no longer in the sale price variable.

We used RMSE(root mean square error) method for model evaluation.

Text

Description automatically generated with medium confidence

A picture containing diagram

Description automatically generated

To evaluate the performance of a model is to train it on a number of different smaller datasets and evaluate them over the other smaller testing set.

1. Linear Regression

Text

Description automatically generated

1. Random Forest Regressor

Graphical user interface, text, application

Description automatically generated

1. Gaussian Process Regressor

Text

Description automatically generated

1. XGBoost Regressor

Text

Description automatically generated

1. Stocking method

Graphical user interface, text, application

Description automatically generated

**Roles and Responsibilities:**

We are a four-person team that initially evaluated the dataset and split our work according to our expertise: Harsh conducted linear regression and preprocessing, Jayraj did Gaussian Process, Sophia handled Random Forest and ensembled model through the stocking approach, and Nidhi worked on EDA as well as XGBoost.

**Conclusion:**

The data was trained using a Random Forest Regressor, which resulted in the lowest RMSE. The next-to-lowest RMSE was attained by the Stacking model, which used Linear Regression as its final classifier.