AI - PS 1

AIM

To develop a machine learning model to predict the sale prices of the houses depending on the various factors given.

DATA PROVIDED

1. **train.csv** - the training set
2. **test.csv** - the test set
3. **data_description.txt** - full description of each column
4. **sample_submission.csv** - a benchmark submission from a linear regression on year and month of sale, lot square footage, and number of bedrooms

DATA PREPROCESSING

1. Libraries Imported: Essential libraries like **pandas**, **numpy**, **matplotlib.pyplot**, and **seaborn** were imported for data manipulation and visualization.
2. Data Loading:
    - The **train.csv** and **test.csv** files were loaded into pandas DataFrames.
    - A new column **Source** was added to distinguish between training and test data, and the datasets were concatenated.
3. Handling Missing Values:
    - Columns with more than 500 null values were dropped.
    - Missing values in numerical columns were filled using the median.

    - Missing values in categorical columns were filled using mode imputation.
4. Categorical Encoding:
    - One-hot encoding was applied to categorical variables to transform them into a format suitable for machine learning models.
    - This method was chosen over integer mapping as it prevents the creation of an ordinal relationship between categories, which could mislead the model.
5. Feature Selection:
    - A correlation matrix was used to identify the most important features correlated with the **SalePrice.**
    - The features were then selected based on their correlation values.
6. Data Splitting:
    - The combined dataset was split back into training and test datasets.
    - The training data was further split into a training set (80%) and a validation set (20%) for model evaluation.

MODEL SELECTION AND TRAINING

- Model Chosen: A stacking ensemble model was used, combining three powerful base models:
    - XGBoost: Known for handling large datasets efficiently.
    - LightGBM: A gradient boosting framework that uses tree-based learning algorithms.

    - CatBoost: A gradient boosting algorithm that handles categorical data without extensive preprocessing.
- Stacking Model: A **StackingRegressor** was used, with Ridge regression as the final estimator.
- Training:
    - The stacking model was trained on the training dataset.
    - After training, the model was evaluated on the validation dataset.

MODEL EVALUATION

- Evaluation Metrics:
    - **Logarithmic RMSE: 0.1457**
    - **R^2 Score: 0.8862**
- These metrics indicate a good fit, with the model explaining approximately 88.6% of the variance in the target variable.

PREDICTIONS ON TEST DATASET

- Final Model Training: The stacking model was retrained on the entire training dataset before predicting on the test set.
- Predictions: Predictions for the test dataset were generated and saved in **test_predictions.csv**.