WEEK 5 ASSIGNMENT

🏠 House Price Prediction with RMSE Minimization and Stacking Ensemble

1. Import Required Libraries

We begin by importing necessary libraries for data manipulation, visualization, preprocessing, and modeling.

Output: Library import successful.

2. Load Datasets

The training and testing datasets, along with the sample submission, are loaded using pandas.

Output:

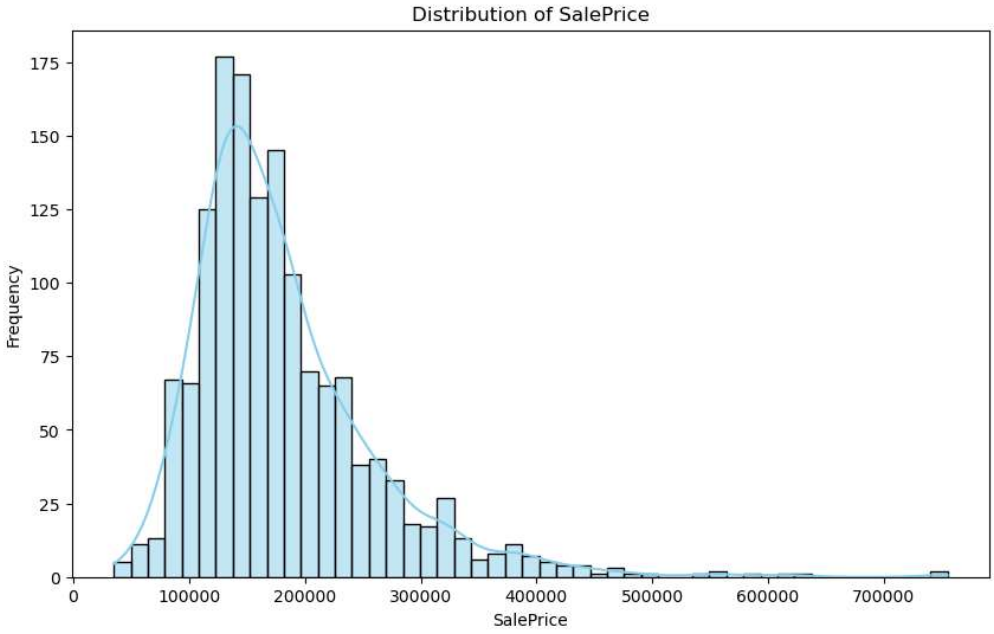
```
Train shape: (1460, 81)
Test shape: (1459, 80)
```

3. Visualize Target Variable Distribution

A histogram is plotted to analyze the distribution of the target variable SalePrice.

Output:

Histogram plot of SalePrice.

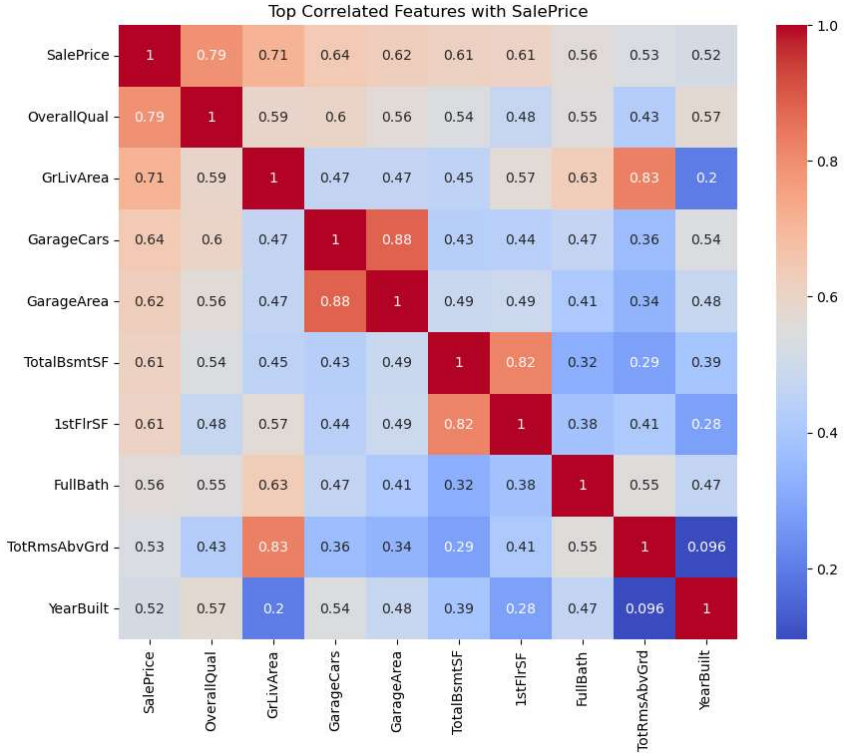


4. Correlation Heatmap of Top Features

We identify the top 10 features most correlated with SalePrice and visualize their correlation.

Output:

Heatmap of top correlated features.



5. Combine Train and Test Data

Combine both datasets into a single DataFrame for uniform preprocessing.

Output: Combined dataset info.

6. Missing Data Heatmap

Visualize missing values using a heatmap.

Output:

Missing data heatmap.

7. Impute Missing Values

All missing values are filled using median for numeric and mode for categorical features.

Output: Missing values handled.

8. Label Encoding for Ordinal Features

Selected ordinal categorical features are encoded using Label Encoding.

Output: Ordinal features encoded.

9. One-Hot Encoding

One-hot encoding is applied to nominal categorical features.

Output:

Shape after encoding.

```
Shape after one-hot encoding: (2919, 253)
```

10. Feature Engineering

Create new useful features such as total area, number of bathrooms, and age-related columns.

Output: Engineered features added.

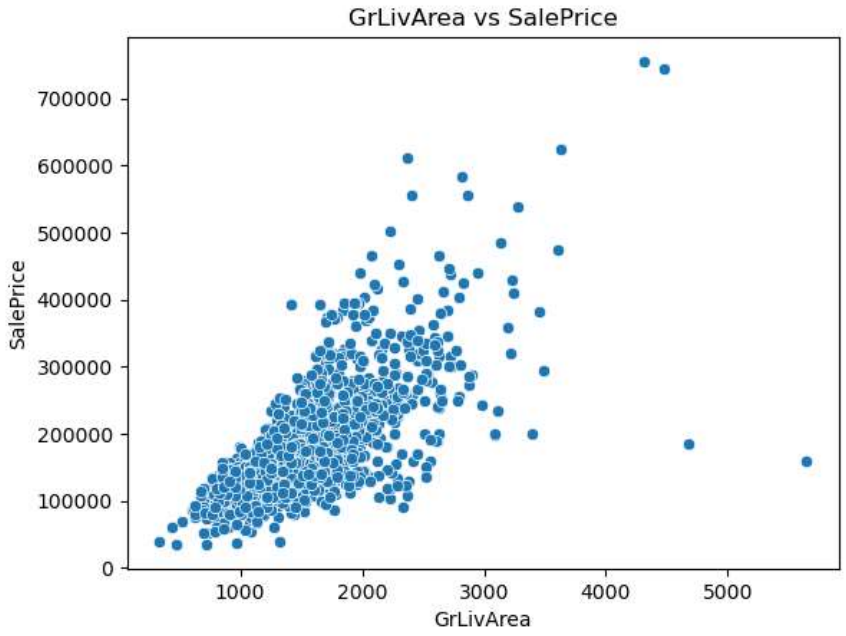
11. Remove Outliers

Detect and remove outliers from GrLivArea vs SalePrice plot.

Output:

Outlier-removed data.

Plot to Remove Outliers



12. Feature Scaling

Standardize features using StandardScaler.

Output: Scaled training and testing sets.

13. Model Evaluation Function

Define RMSE calculation function using 5-fold cross-validation.

Output: RMSE function ready.

14. Base Model Training

Train base models: Random Forest and Gradient Boosting on log-transformed target.

Output: Models trained successfully.

15. Evaluate Base Models

Evaluate models using RMSE on log-transformed target.

Output:

```
RF CV RMSE: 0.14258522248098918
GB CV RMSE: 0.1329387457414125
```

16. Stacking Ensemble (RF + GB)

Use stacking ensemble with Random Forest and Gradient Boosting as base models and Linear Regression as the final estimator.

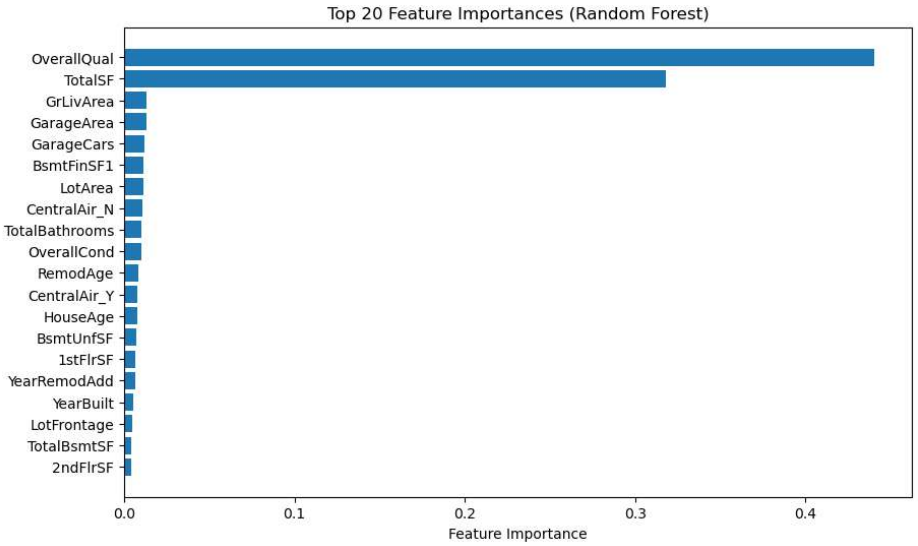
Output: Stacked model trained and predictions generated.

17. Feature Importance (Random Forest)

Visualize top 20 most important features from the Random Forest model.

Output:

Bar plot of feature importances.



18. Submission (Using Stacking Model)

Generate final predictions using stacked model and save them to submission.csv.

Output: Final CSV file created.

```
Submission saved to predictions.csv
```