House Pricing Prediction

Team 70

Youssef Khaled Abdallah *20191700767*

Youssef Omar Hussein *20191700776*

Youssef Khaled Farouk *20191700768*

Required Libraries

1. **Caret**.
2. **RandomForest**.
3. **GBM**.

Preprocessing

1. `ExploreData(data)`

   displays the number of columns, number of rows, count of missing values, and summary statistics of the data

2. `RemoveNullColumns(threshold, train_data, test_data)`

   Drop columns that have more than a **threshold** percentage of **NULL** values. It calculates the **NULL** percentage for each column, identifies columns exceeding the threshold, and removes those columns from the train and test datasets.

3. `EncodeCategoricalData(train_data, test_data)`

   Performs **one-hot encoding** on **categorical variables** in the train and test datasets. It saves the target variable from the train dataset, removes the target variable, combines train and test datasets, applies **one-hot encoding**, and **normalizes** the data using **Z-scores**. It splits the encoded data back into train and test datasets.

4. `FillNulls(data)`

   Replace **NULL** values in a dataset with the median value of each respective column.

5. `Preprocess(train_data, test_data)`

   Integrate the preprocessing steps by calling the **RemoveNullColumns**, **EncodeCategoricalData**, and **FillNulls** functions. It returns the preprocessed **x_train, x_test, y_train, y_test, train_ids, and test_ids.**

6. `TrainValidateSplit(x, y, ids, splitSize=0.8, seed=100)`

   Shuffles and splits the preprocessed train dataset into training and validation sets. It adds IDs to the datasets, randomly partitions the data based on a specified split size and returns the split datasets.

7. `RemoveLowCorrelation(x_train, y_train, threshold)`

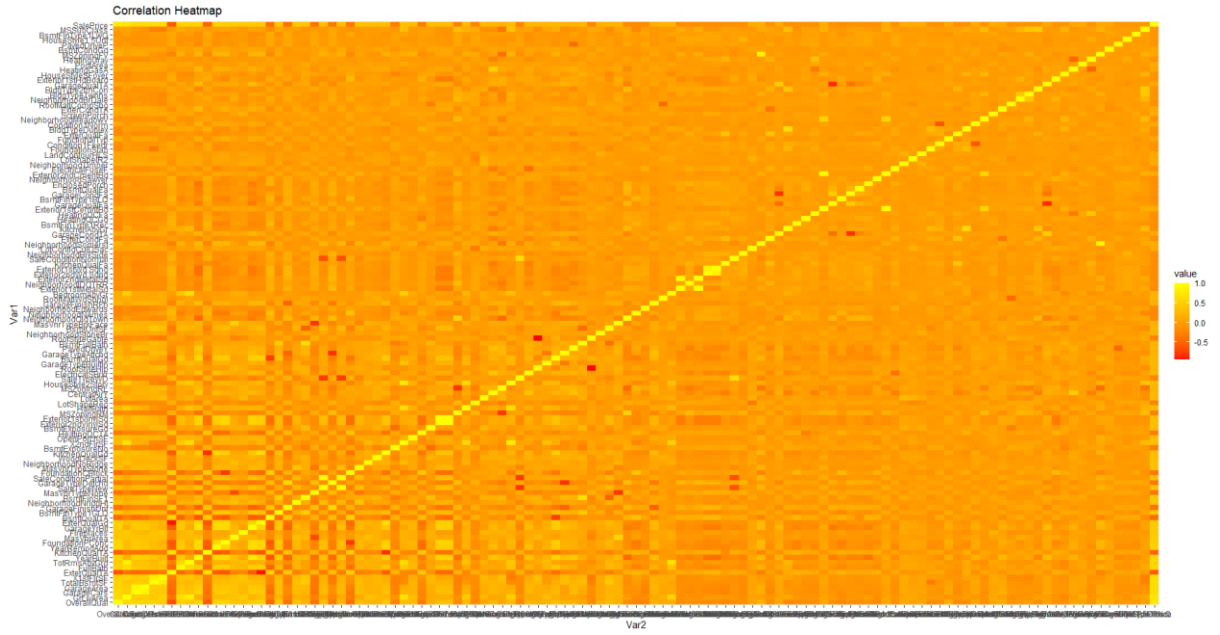
   Removes Least **threshold** percentage correlation columns

Model

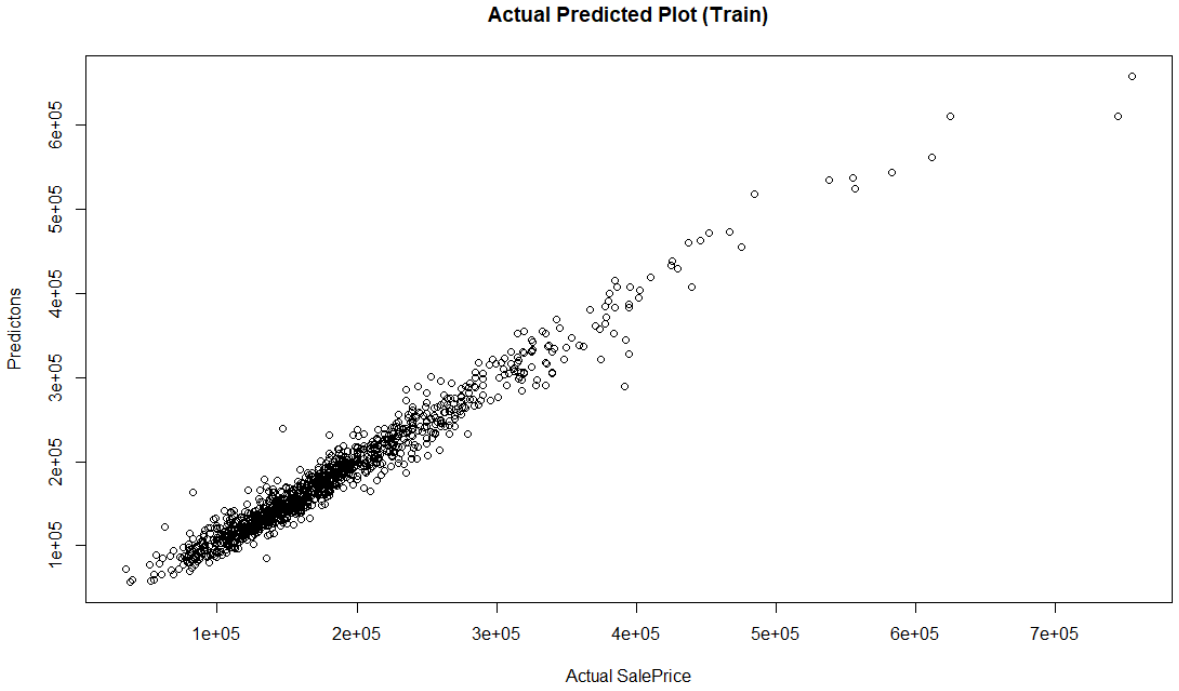
Gradient Boosting machine : It is a type of ensemble method that combines multiple weak prediction models (typically decision trees) to create a strong predictive model.

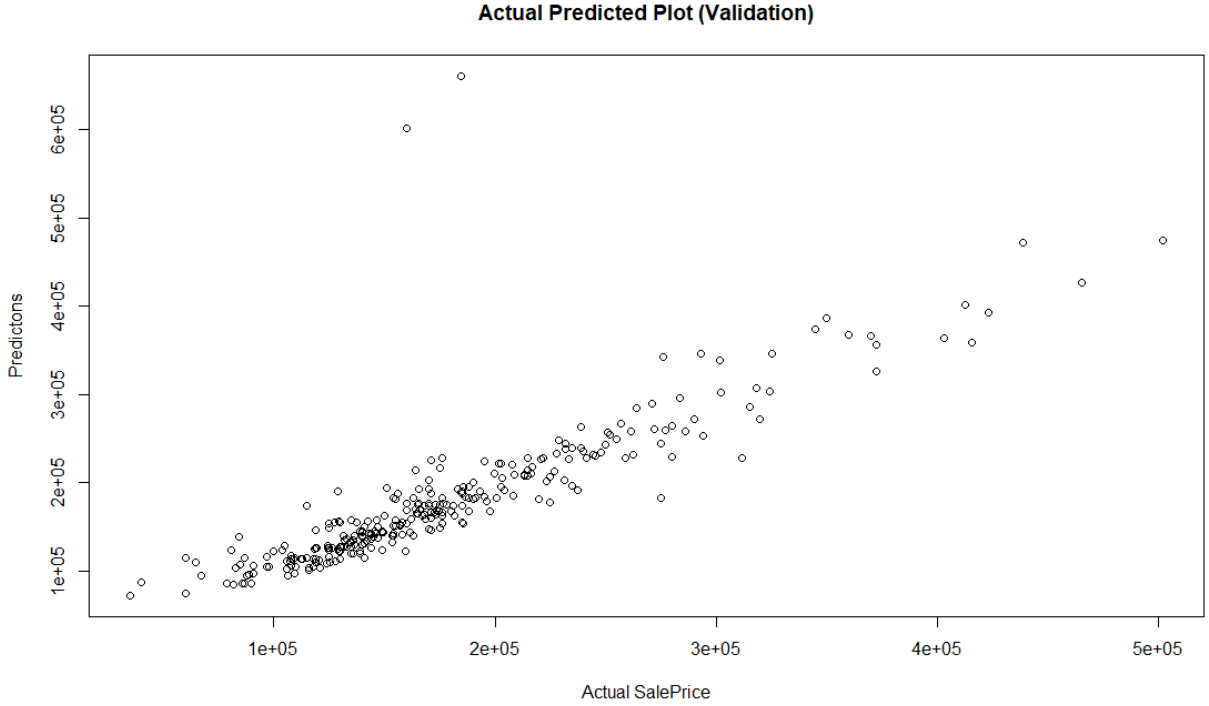
Data Visualization

Variables Correlations Heatmap

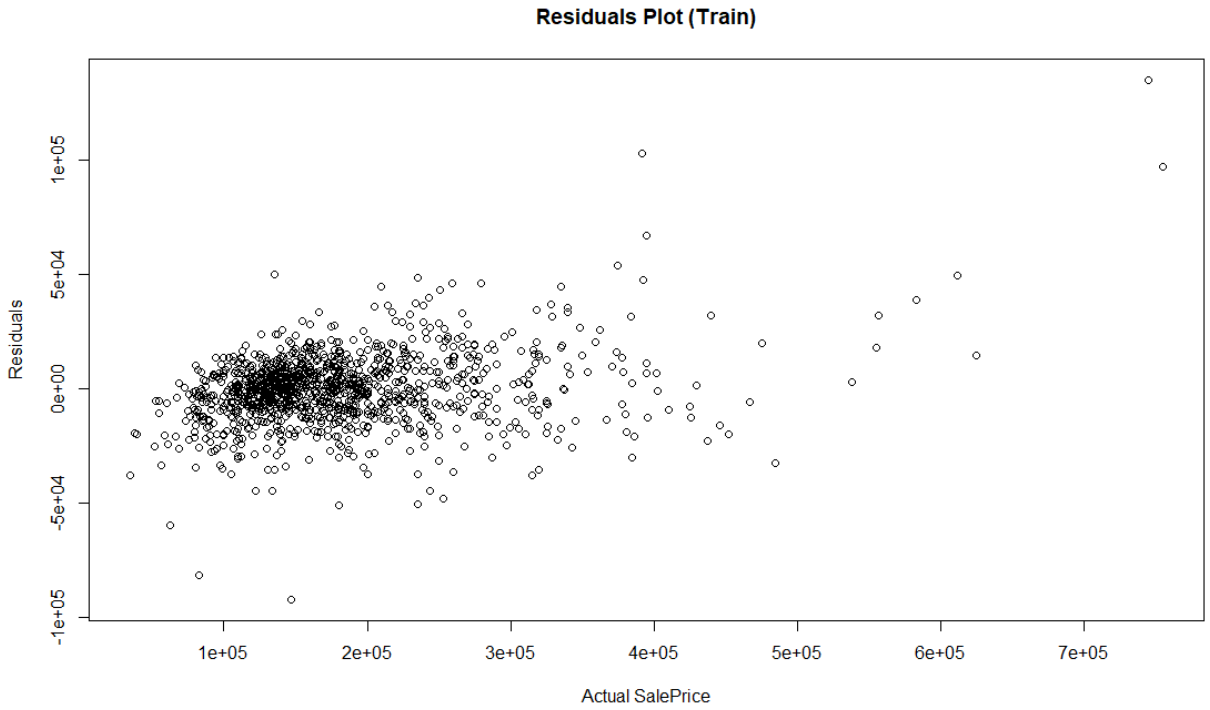


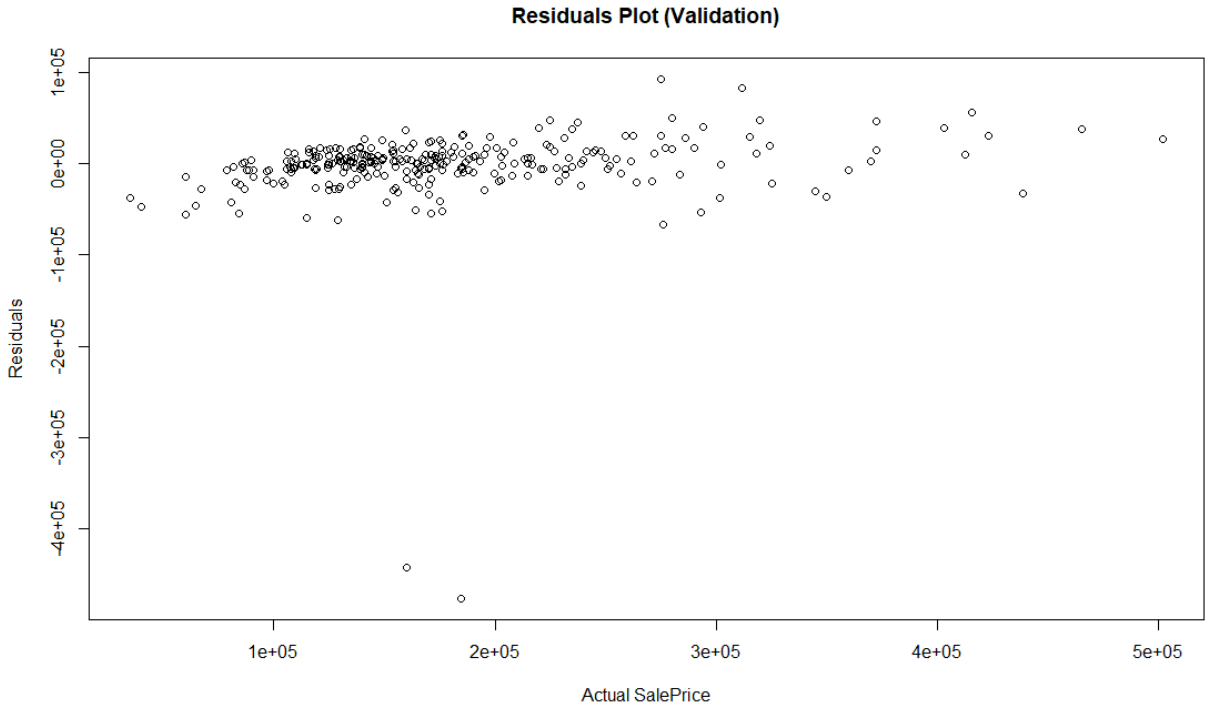
Actual vs Predicted Plot:



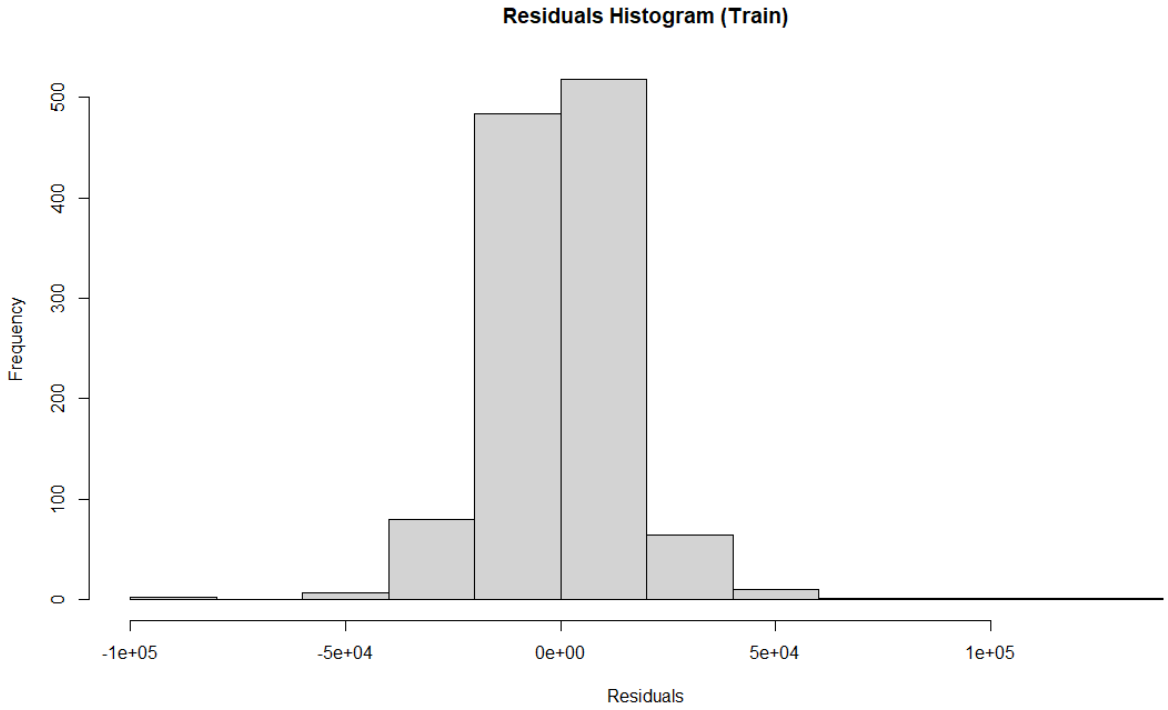


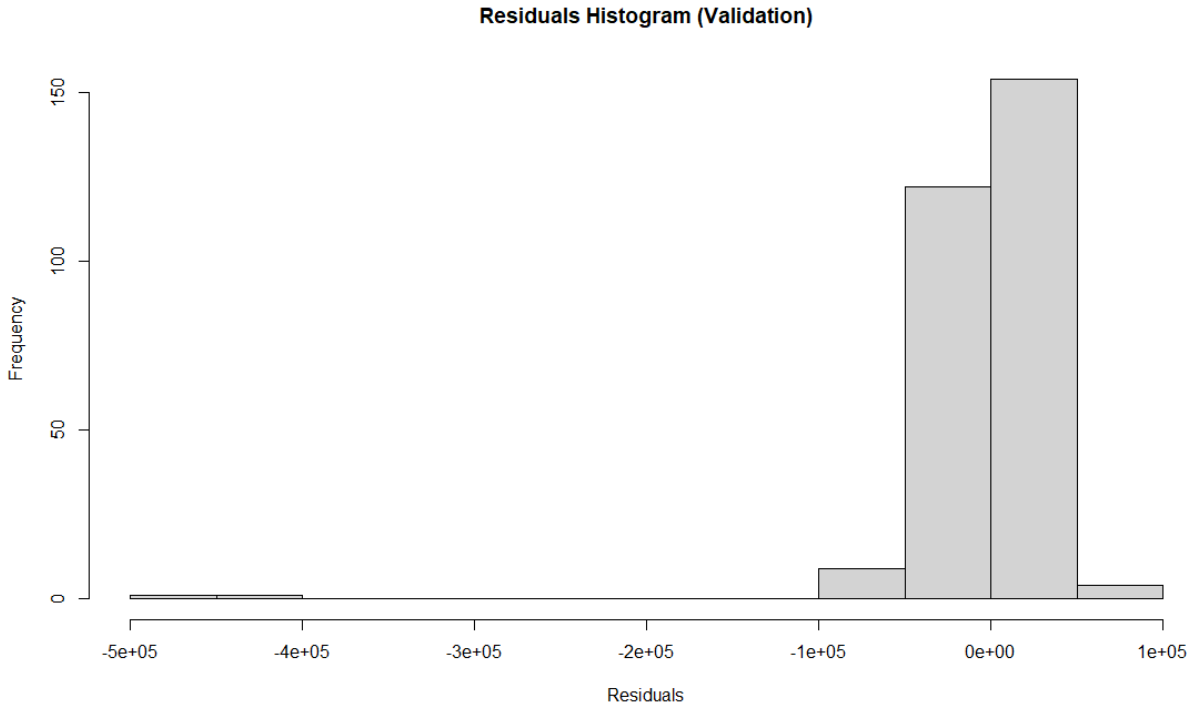
Residuals Plot:





Residuals Histogram:





Evaluation Metrics

The evaluation metrics for the **training dataset** are as follows:

- Mean Squared Error (MSE): 249,788,746
- Root Mean Squared Error (RMSE): 15,804.71
- R-squared (R^2): 0.9619353

The evaluation metrics for the **validation dataset** are as follows:

- Mean Squared Error (MSE): 1,912,787,428
- Root Mean Squared Error (RMSE): 43,735.43
- R-squared (R^2): 0.710436

Summary

We have successfully applied preprocessing techniques, data visualization, utilized the Gradient Boosting Machine model, and evaluated the performance of house pricing prediction model using various metrics.