



PREDICTING HOUSING PRICE



Introduction:

This report presents an analysis of house price prediction using various models. The objective of this analysis is to develop accurate models that can effectively predict house prices based on given features. Accurate house price prediction is crucial for real estate professionals, homeowners, and potential buyers, as it provides valuable insights for decision-making processes such as property valuation, investment strategies, and pricing negotiations.

Preprocessing:

In the preprocessing step, we performed the following tasks on the dataset:

1. Removed unnecessary columns: We removed the columns "Street," "Alley," "Utilities," "PoolQC," "Fence," "MiscFeature," and "FireplaceQu" from both the training and test datasets.
2. Imputed missing values: We imputed missing values in the dataset using the ``impute_missing`` function. For numeric columns, we replaced missing values with the column mean, and for categorical columns, we replaced missing values with the mode.
3. Converted data to factors: We converted selected columns to factor type using the ``convert_to_factor`` function. This was done for columns such as "MSZoning," "HouseStyle," "LotShape," and many others.

4. Converted data to numeric: We converted selected columns to numeric type using the ``convert_to_numeric`` function. This was done for the same set of columns that were converted to factors.

Models Used:

We used the following models for house price prediction:

1. Linear Regression: We trained a linear regression model using the ``lm`` function with the formula "SalePrice ~ ." This model considers all available features in the dataset.

2. Support Vector Machine (SVM): We trained an SVM model using the ``svm`` function from the "e1071" package. Again, we used the formula "SalePrice ~ ." to consider all features.

3. Random Forest: We trained a random forest model using the ``randomForest`` function from the "randomForest" package. Similar to the previous models, we used the formula "SalePrice ~ ." to include all features.

Performance:

<div>model</div> <div>evaluation</div>	Linear Regression	Support Vector Machine	Random Forest
MSE	1187555247	311651241.04	83802472
RMSE	34460.92	17653.65	9154.37
R2	0.81	0.95	0.99

Visualization:

We created a scatter plot titled "Actual vs Predicted SalePrice for all the models (Linear Regression, SVM, Random Forest)

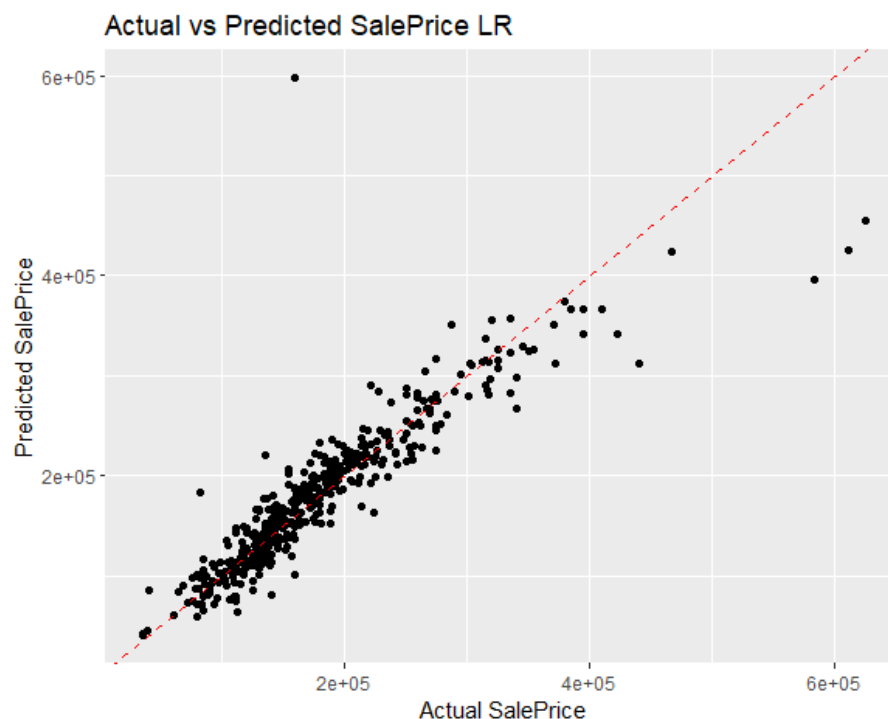
The scatter plots allow us to assess the performance of the models visually. The plots show how well the models predicted the sale prices compared to the actual values.

We created two bar plots of the RMSE and R-squared values for three different models: linear regression, support vector machine (SVM), and random forest (RF). Both bar plots have three bars, one for each model, and the x-axis shows the different model names.

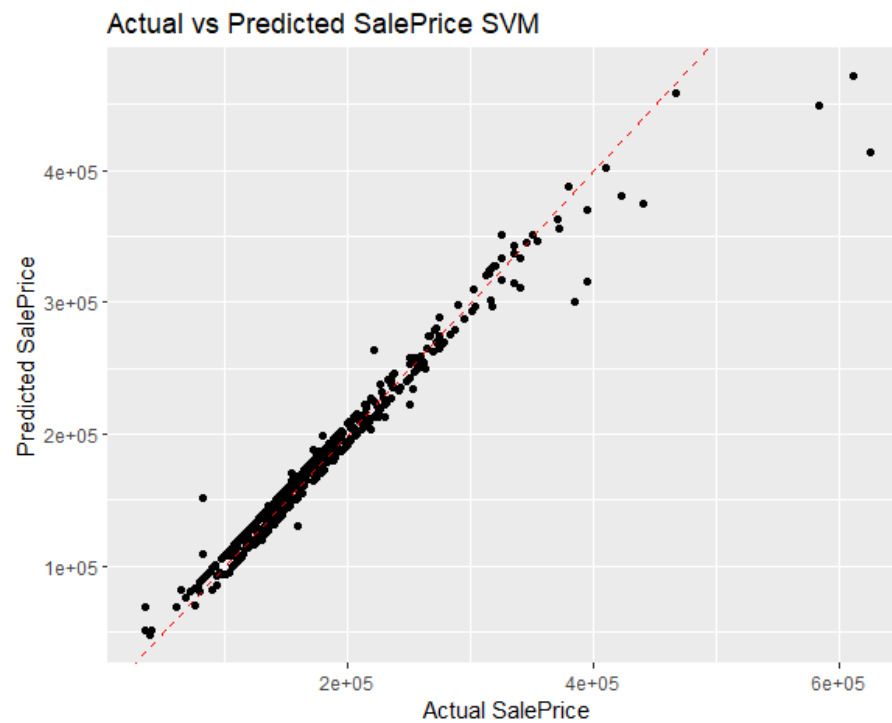
The bar plots provide a visual comparison of the performance of the three models based on their RMSE and R-squared values.

Plots:

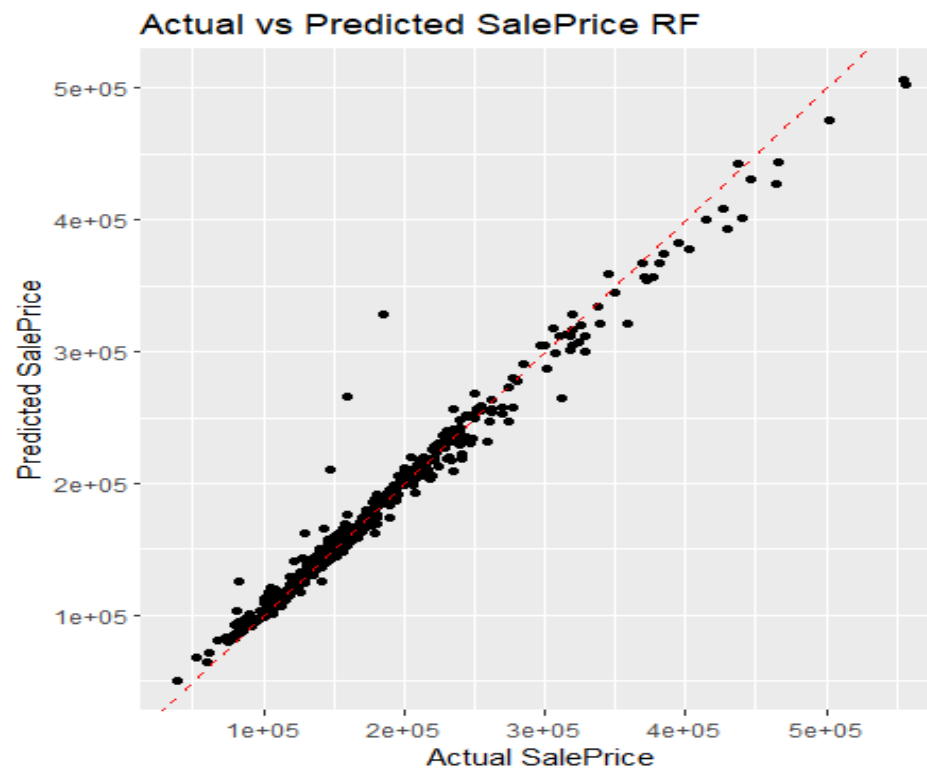
1- Linear Regression:



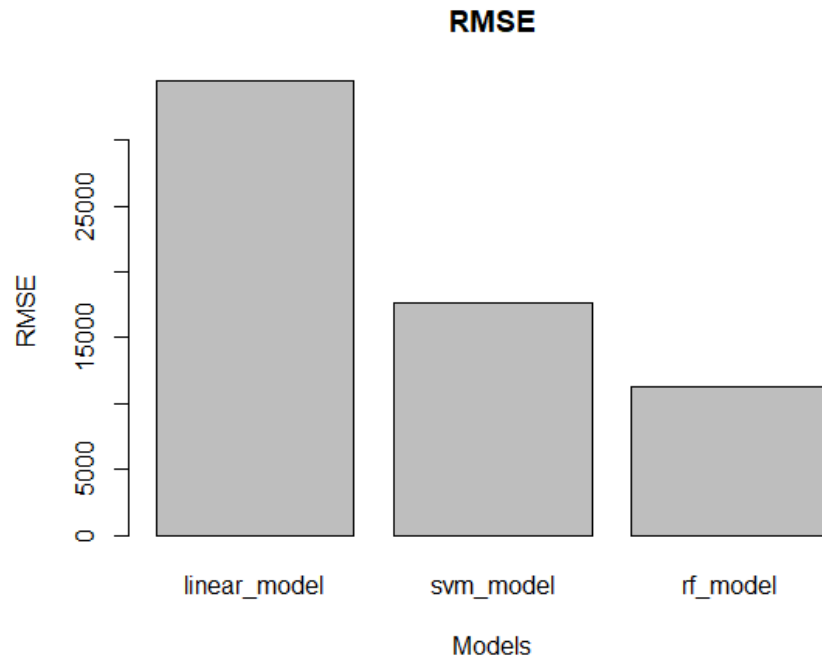
2-Support Vector Machine(SVM):



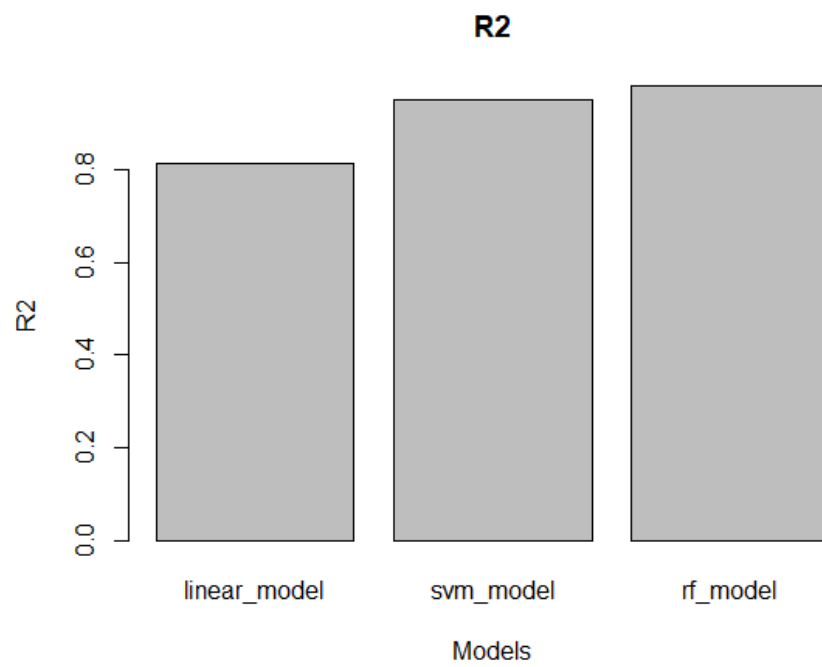
3- Random Forest:



Bar plots of the RMSE:



Bar plots of the R-squared :



Conclusion:

In this analysis, we performed preprocessing on the house price dataset by removing unnecessary columns, imputing missing values, and converting data types. We then used three different models, including linear regression, SVM, and random forest, to predict house prices.

The models were evaluated using metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared (R^2). Based on these metrics, we observed that the random forest model achieved the lowest RMSE and highest R^2 , indicating better performance compared to the other models.

The scatter plots provided visual representations of the predicted sale prices against the actual sale prices, allowing us to assess the model predictions visually.

Overall, the random forest model showed the most promising results for predicting house prices in this analysis.