National University of Computer and Emerging Sciences

House Price Prediction

From
Moin Sultan (19I-0934)
Safwan Siddique (20I-0982)
Arham Asjid (20I-2193)

For
Mr. Hamad Ul Qudous

13th May, 2024

Introduction

House is a place where everyone feels relaxed. It is a place where a person takes a sigh of relief. It is a place where everyone wants to return back after being tired working all day, but not all the people have their own house. Some people live on rent, some in tents and some on the streets. Despite all of this, everyone wants to have their own house. So this project is to assist those people who are looking to buy houses.

Problem Statement

Everyone wants to have their own house in a fully developed area, with all infrastructure and other facilities, with a certain amount of area, a certain number of bedrooms, a certain number of bathrooms, and a lot of other requirements as well. When the customer gives this long list of requirements to a property dealer, the dealer becomes exhausted and is unable to find a match that satisfies all these requirements and thus cannot provide the customer with an estimate of how much the price of a house with such requirements would be. So our aim is to build a system that predicts the price of the house based upon certain requirements which can assist the customers as well as the property dealers.

Methodology

To solve the problem we have used the concepts of Machine Learning. We have trained a machine learning model on a huge dataset, related to house prices, which consists of 21613 instances spreaded over 21 different features like: Lot Area (in sqft), no. of bedrooms, no. of bathrooms, no. of floors etc. The model takes the customer requirements regarding all the 21 different features for a house, and predicts a price for it. To train the model we have gone through a number of steps like: Data Preprocessing, Exploratory Data Analysis, Feature Extraction, Model Training and Validation, Model Testing and Model Evaluation.

- **Data Preprocessing**
  To make data suitable for training the machine learning model we have taken a number of preprocessing steps that are as follows:
  - Null Value Handling
  - Outlier Removal
  - Binary Encoding (for ‘Waterfront View’ feature)
  - One Hot Encoding (for ‘No of Times Visited’ and ‘Condition of the House’ features)
  - Normalization (Standard Scaler)

- **Exploratory Data Analysis**
  To explore the dataset and find the deeper insights about it we have used a number visualizations like:
  - Histogram
  - Scatterplot
  - Barplot
  - Correlation Matrix

The visualizations of this part are added in the “Visualization outputs” section.

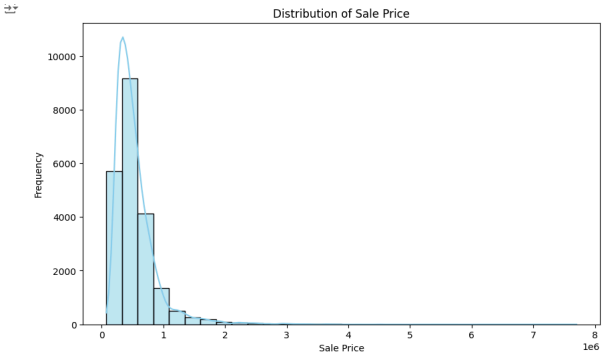
- **Model Training and Validation**
  To train the model our target variable was ‘Sale Price’ and we used the rest of the features as independent variables except ‘Date House was Sold’. We splitted the data into train, validation and test sets with 60%, 20% and 20% data of the original dataset, respectively. As we were to predict a numeric value so this represents that it is a regression problem. So to solve this regression problem we used 3 regression models: Linear Regression, Ridge Regression and Random Forest Regressor. All these models were trained on the training dataset and then validated on the validation dataset.

- **Model Testing**
  After training and validating the models, we moved on to the testing part. The model was tested on the test dataset and the results were evaluated.

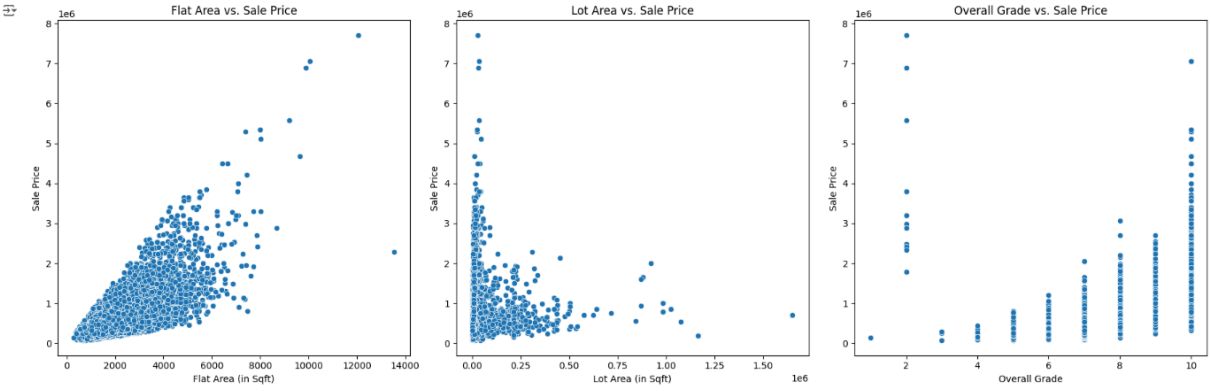
- **Model Evaluation**
  To evaluate the model we used the evaluation metrics for the regression problem like: Mean Squared Error (MSE), Mean Absolute Error and R Squared Error. Based on these evaluation metrics, the regression models were analyzed to identify the best one.

Visualization Outputs

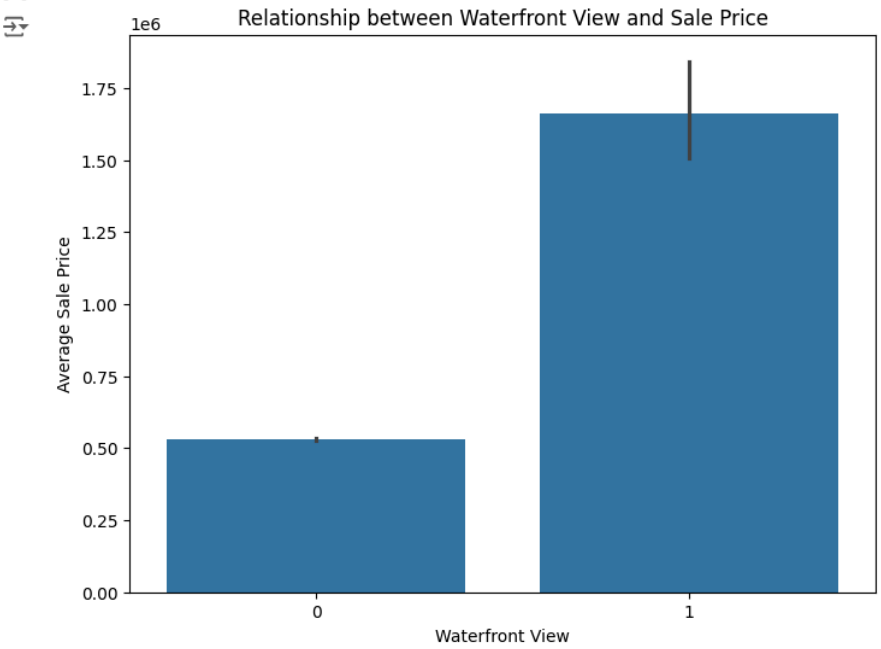
1. Histogram



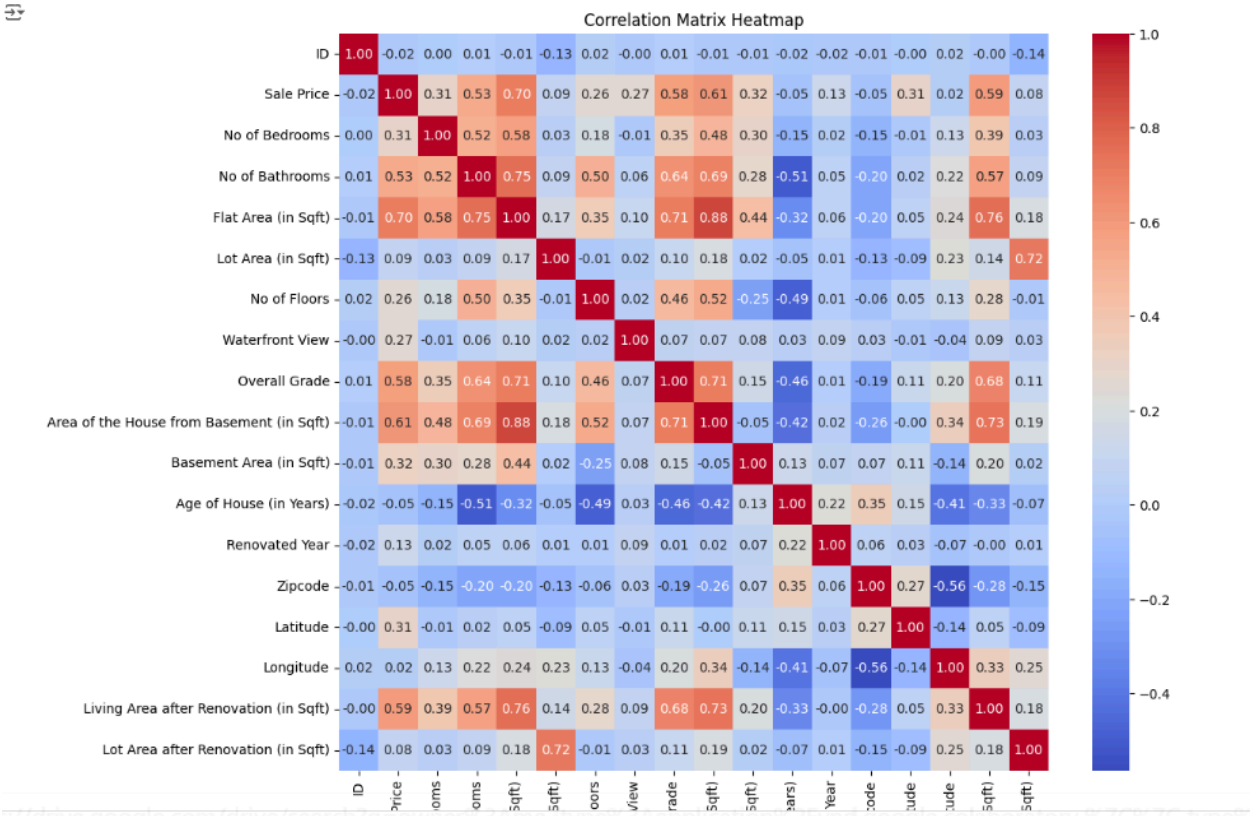
2. Scatterplot



3. Barplot



4. Correlation Matrix



Results

```
test_and_evaluate_models(X_test, y_test, models, 'Test');
```

```
Linear Regression: Test MSE = 45582057472.0
Linear Regression: Test RMSE =213499.546875
Linear Regression: Test MAE =125539.28125
Linear Regression: Test R2 =0.6987929676257162


Ridge Regression: Test MSE = 45579476992.0
Ridge Regression: Test RMSE =213493.5
Ridge Regression: Test MAE =125509.671875
Ridge Regression: Test R2 =0.6988100157200289


Random Forest Regression: Test MSE = 22410289160.509346
Random Forest Regression: Test RMSE =149700.6651972841
Random Forest Regression: Test MAE =75889.6021196098
Random Forest Regression: Test R2 =0.8519124237588291
```

Based on the results that are being produced after applying the evaluation metrics, Random Forest Regression proves to be the best model for this problem because its performance is best in comparison to Linear and Ridge Regression as justified by the R2_Score.