

MBA Semester – IV

Research Project – Interim Report

|  |  |
| --- | --- |
| **Name** | Jaffer Sadhik |
| **Project** | House price prediction |
| **Group** | Data science and analytics |
| **Date of Submission** | 27 Apr 2024 |



**A study on “House Price Prediction”**

## Research Project submitted to Jain Online (Deemed-to-be University)

## In partial fulfillment of the requirements for the award of:

**Master of Business Administration**

*Submitted by:*

**Jaffer Sadhik**

USN:

**221VMBR01121**

*Under the guidance of:*

**Dr. C. S. Jyothirmayee**

(Faculty-JAIN Online)

Jain Online (Deemed-to-be University)

Bangalore

**2022-23**

**DECLARATION**

I, Jaffer Sadhik, hereby declare that the Research Project Report titled “House Price Prediction” has been prepared by me under the guidance of the Dr. C. S. Jyothirmayee. I declare that this Project work is towards the partial fulfillment of the University Regulations for the award of the degree of Master of Business Administration by Jain University, Bengaluru. I have undergone a project for a period of Eight Weeks. I further declare that this Project is based on the original study undertaken by me and has not been submitted for the award of any degree/diploma from any other University / Institution.

Place: Bangalore \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: 27 Apr 2024 Jaffer Sadhik

221VMBR01121

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **Title** | **Page Nos.** |
| List of Tables |  |
| List of Graphs |  |
| Executive Summary | 1 |
| Chapter 1: Introduction and Background | 2-6 |
| Chapter 2: Research Methodology | 8-15 |
| Chapter 3: Data Analysis and Interpretation | 17-18 |
| References | 19 |

|  |  |
| --- | --- |
| **List of Graphs** | |
| **Graph Title** | **Page No.** |
| Bar graph | 10-11 |
| Box Plot | 11 |
| Scatter Plot | 12 |
| Matrix Plot | 13 |
| Heat Matrix | 17 |

**CHAPTER 1**

**INTRODUCTION AND BACKGROUND**

1. **INTRODUCTION AND BACKGROUND**

**1.1 Executive Summary**

In the real estate industry, accurately predicting house prices is important for making informed decisions. This project focuses on using machine learning and data analysis techniques to predict house prices effectively. The process begins with compiling a comprehensive dataset containing property details such as size, location, bedrooms, bathrooms, etc. Data cleaning and preparation is done to ensure the dataset is free from anomalies and missing values. Then feature engineering is done to enhance prediction accuracy by creating additional variables.

Model training involves dividing the dataset into training and testing sets, allowing the model to learn the relationships between input features and target pricing. Performance evaluation metrics such as mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) are used to assess model performance, with cross-validation techniques ensuring model resilience.

Upon model development, it can be integrated into real estate applications or websites to provide quick property value estimates. However, predicting house prices accurately is challenging due to numerous influencing variables and varying property conditions.

To address this challenge, the project aims to analyze available property data using Exploratory Data Analysis (EDA) to predict prices accurately, benefiting both buyers and sellers.

Exploratory Data Analysis (EDA) plays an important role in understanding the dataset, identifying patterns, abnormalities, Co-relations, and confirming that the right questions are being asked and analyzed. Python, along with libraries like Pandas, Numpy, Matplotlib, and Seaborn, will be utilized for EDA by providing tools for data visualization and statistical analysis. Depending on the number of columns we are analyzing we can divide EDA into two types:

1. Univariate Analysis – In univariate analysis, we analyze or deal with only one variable at a time. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.
2. Bi-Variate analysis – This type of data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship between the two variables.

The project involves conducting EDA on the dataset, including descriptive statistics, identifying missing values, duplicates, abnormalities, and visualizing univariate and bivariate relationships. Various predictive models, such as linear regression, are applied to the dataset, and model evaluation metrics are analyzed to identify the best fit.

Conducting EDA helps organize the dataset before modeling, understand variables and relationships, choose the right model, and identify patterns. Overall, the project aims to provide house buying and selling guiding services, ensuring fair pricing and preventing loss for buyers and sellers alike.

**1.2 Introduction and Background**

The real estate market plays an important role in the economy and people's lives. Owning a house is a major financial decision for many individuals and families. It is also considered as an asset for those seeking to use it as an investment opportunity.

Predicting house prices is a complex task that involves utilizing various data sources, statistical techniques, and machine learning algorithms. These predictive models can provide valuable assistance and insights to different stakeholders:

* Homebuyers: To make informed decisions by gaining insights into market trends.
* Real Estate Professionals: To optimize their portfolios and provide better guidance to clients.
* Investors: To Identify potential investment opportunities.

Factors influencing house prices are numerous and complex, ranging from location and property size to economic trends, interest rates, and local amenities. Predicting house prices accurately is not only challenging but also highly valuable.

Recent advancements in data analytics and machine learning have transformed the field of house price prediction. Data scientists and real estate professionals now have access to vast amounts of data, enabling them to develop sophisticated predictive models that can navigate the evolving landscape of the real estate market.

**1.3 Problem Statement**

Accurately estimating a house's worth in the real estate market goes beyond just location and size. A house's price is influenced by a multitude of complex factors. Homeowners often face the challenge of setting a competitive asking price that is fair yet not so high that it discourages potential buyers.

Traditionally, people have relied on comparisons with similar properties in the neighborhood to determine a house's value. However, this method can be unreliable and subjective.

**1.4 Objective of the Study**

This study aims to achieve the following objectives:

* To develop a robust price prediction model for houses.
* To identify the key house price attributes that significantly influence the model's predictive power.
* To validate the model's accuracy in predicting house prices.

We will be utilizing the provided dataset containing various features to analyze and predict house prices. The features include:

* cid: Unique identifier for the house within the dataset.
* day\_hours: Date when the house was sold.
* price: The price is the prediction target.
* room\_bed: Number of bedrooms in the house.
* room\_bath: Number of bathrooms per bedroom.
* living\_measure: Square footage of the home.
* lot\_measure: Square footage of the lot.
* ceil: Total number of floors (levels) in the house.
* coast: Indicates if the house has a view of the waterfront.
* sight: Indicates if the house has been viewed.
* condition: The overall condition of the house.
* quality: Grade given to the housing unit, based on the grading system.
* ceil\_measure: Square footage of the house excluding the basement.
* basement\_measure: Square footage of the basement.
* yr\_built: Year the house was built.
* yr\_renovated: Year the house was renovated.
* zip\_code: ZIP code.
* lat: Latitude coordinate.
* long: Longitude coordinate.
* living\_measure15: Living room area in 2015 (implies some renovations), which may or may not have affected the lot size area.
* lot\_measure15: Lot size area in 2015 (implies some renovations).
* furnished: Indicates the furnishing quality of the room.
* total\_area: Combined area of the living space and the lot.

**1.5 Company and Industry Overview**

The real estate market is highly competitive in terms of pricing, with significant variations based on numerous factors. Prediction of property prices is an important factor for decision-making by both buyers and investors. It can support budget allocation, property search strategies, and policy determination. Therefore, applying machine learning concepts becomes necessary for optimizing and predicting prices with high accuracy in this field.

Investors rely on market trends to maximize returns. Developers need to understand future trends to make informed decisions about project feasibility and potential challenges. To accurately estimate property prices and future trends, a large amount of data that influences land prices is required for analysis, modeling, and predicting.

* 1. **Overview of Theoretical Concepts**
* Data Collection and Preprocessing: Gathering relevant data, cleaning inconsistencies, and preparing it for analysis.
* Exploratory Data Analysis (EDA): Understanding data distributions, relationships between variables, and identifying patterns through visualization and statistical techniques.
* Feature Engineering: Selecting the most relevant features and potentially creating new ones to improve model performance.
* Machine Learning Models: Exploring different regression algorithms (e.g., linear regression, random forests) suitable for price prediction tasks.
* Model Evaluation and Selection: Assessing model performance using techniques like cross-validation and choosing the best performing model based on appropriate metrics.

**CHAPTER 2**

**Research Methodology**

**RESEARCH METHODOLOGY**

**2.1 Scope of the Study**

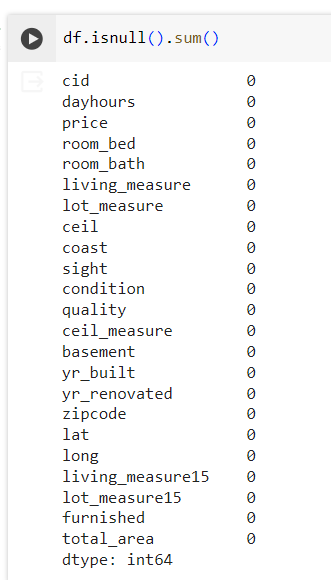
This study delves into the real estate market, aiming to provide a comprehensive analysis through price predictions, data-driven insights, and valuable recommendations. By utilizing the available feature variables and employing powerful predictive modeling techniques, we can empower various stakeholders to make informed decisions regarding property purchases, sales, and listings.

* 1. **Methodology**

**2.2.1 Data Acquisition and Preprocessing**

* **Data Source:** in this project “innercity.xlsx” a pre-shared dataset containing observations on houses was used, containing 23 different features relevant to house price prediction.
* **Software used:**
  + Programming Language: Python
  + Development Environment: Google Colab
  + Operating System: Windows 10
* **Libraries:** Several Python libraries were employed to streamline the machine learning model development process. These include: NumPy, Pandas, matplotlib, Seaborn, KNNImputer.etc

**2.2.2 Data Cleaning and Handling Missing Values:**

* + A screenshot of a computer

    Description automatically generatedThe initial data exploration revealed missing values in certain features. To address this, various imputation techniques were employed, such as filling missing values with the mean, median, or most frequent values within the dataset.

**2.2.3 Exploratory Data Analysis (EDA)**

A brief exploration of the data was conducted to gain insights into the distribution of features, identify potential relationships between variables, and uncover any anomalies. This analysis involved various visualization techniques, including:

* A graph with a number of bars

  Description automatically generated**Univariate Analysis:** Examining the distribution of individual features (e.g., histograms, boxplots) to understand central tendency, spread, and potential outliers.

A graph of a bed

Description automatically generatedA graph with a number of blue bars

Description automatically generated

A graph with blue squares

Description automatically generatedA graph with a bar

Description automatically generated

A graph with a bar

Description automatically generated with medium confidence

A blue bar graph with white squares

Description automatically generatedA graph of a distribution of living

Description automatically generatedA graph with a bar

Description automatically generated

A graph with blue bars

Description automatically generatedA graph of a distribution of cell measure

Description automatically generatedA graph with a bar

Description automatically generated

A graph of a distribution of zipcode

Description automatically generatedA graph of quality distribution

Description automatically generated with medium confidence

A graph of a box plot

Description automatically generatedA graph of a distribution of room bath

Description automatically generated

A graph with lines and numbers

Description automatically generatedA diagram of a bath

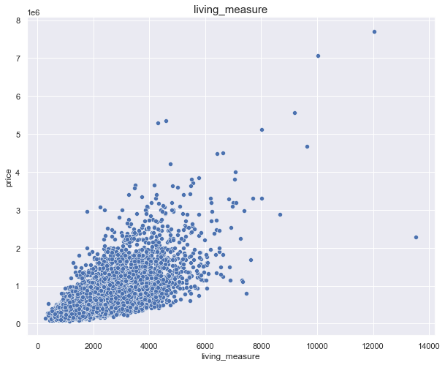
Description automatically generated

A diagram of a bed

Description automatically generatedA diagram of a box plot

Description automatically generated

* **Bivariate Analysis:** Investigating the relationships between pairs of features (e.g., scatter plots) to identify potential correlations or dependencies.

A graph of blue dots

Description automatically generatedA graph showing a number of blue dots

Description automatically generated

A screen shot of a graph

Description automatically generatedA graph of blue dots

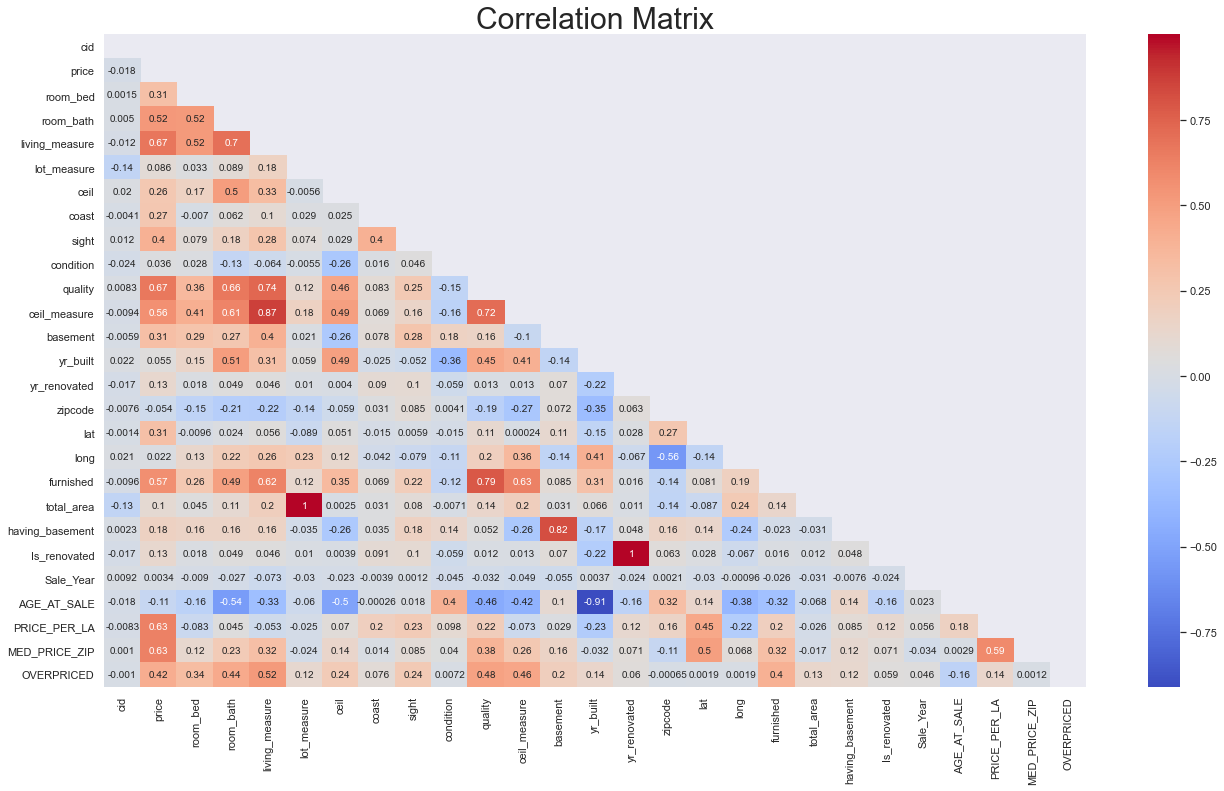
Description automatically generatedA graph showing a number of blue dots

Description automatically generated

A graph showing a number of blue dots

Description automatically generatedA graph showing a number of blue dots

Description automatically generated

* **Multivariate Analysis:** Exploring relationships between multiple features simultaneously (e.g., heatmaps) to gain a more holistic understanding of the data structure.

Through EDA, the identification of features with significant right skewness, categorical features, and potential linear relationships between certain features like price, living area was performed.

**2.2.4 Feature Engineering**

Based on the findings from the EDA, the next step would be to perform feature engineering to improve the performance of the machine learning model by:

* Feature Selection: Choosing a subset of the most relevant features that contribute significantly to the prediction of house prices. Techniques like correlation analysis and feature importance scores can be used for selection.
* Feature Transformation: Addressing skewed distributions by log transformation or creating new features that might be more informative for the model like combining living area and basement area, latitude and longitude.
* Encoding Categorical Features: Converting categorical features into numerical representations suitable for machine learning algorithms.

**2.2.5 Model Building and Evaluation**

The next step is the employment various regression algorithms to develop machine learning models for predicting house prices:

* Linear Regression: A basic yet effective algorithm for modeling linear relationships.
* Decision Trees: Flexible models capable of capturing non-linear relationships between features and the target variable.
* Random Forests: Ensemble learning methods that combine multiple decision trees for improved accuracy and robustness.

The performance of each model would be evaluated using metrics such as mean squared error (MSE) or R-squared. Cross-validation techniques to ensure the generalizability of the models and avoid overfitting.

**2.3 Period of the Study**

The Period of this project is eight Weeks.

**2.4 Utility of the Research**

The findings of this study hold value for various stakeholders in the real estate market:

* Homebuyers: Gaining insights into market trends and price predictions can empower homebuyers to make informed decisions and potentially secure better deals.
* Real Estate Professionals: Data-driven insights can assist realtors in optimizing their strategies, providing more accurate valuations, and guiding clients effectively.
* Investors: The ability to predict price trends can be instrumental for investors seeking to make sound investment decisions in the real estate market.

By providing a deeper understanding of the factors influencing house prices, this research contributes valuable knowledge to the real estate market.

**CHAPTER 3**

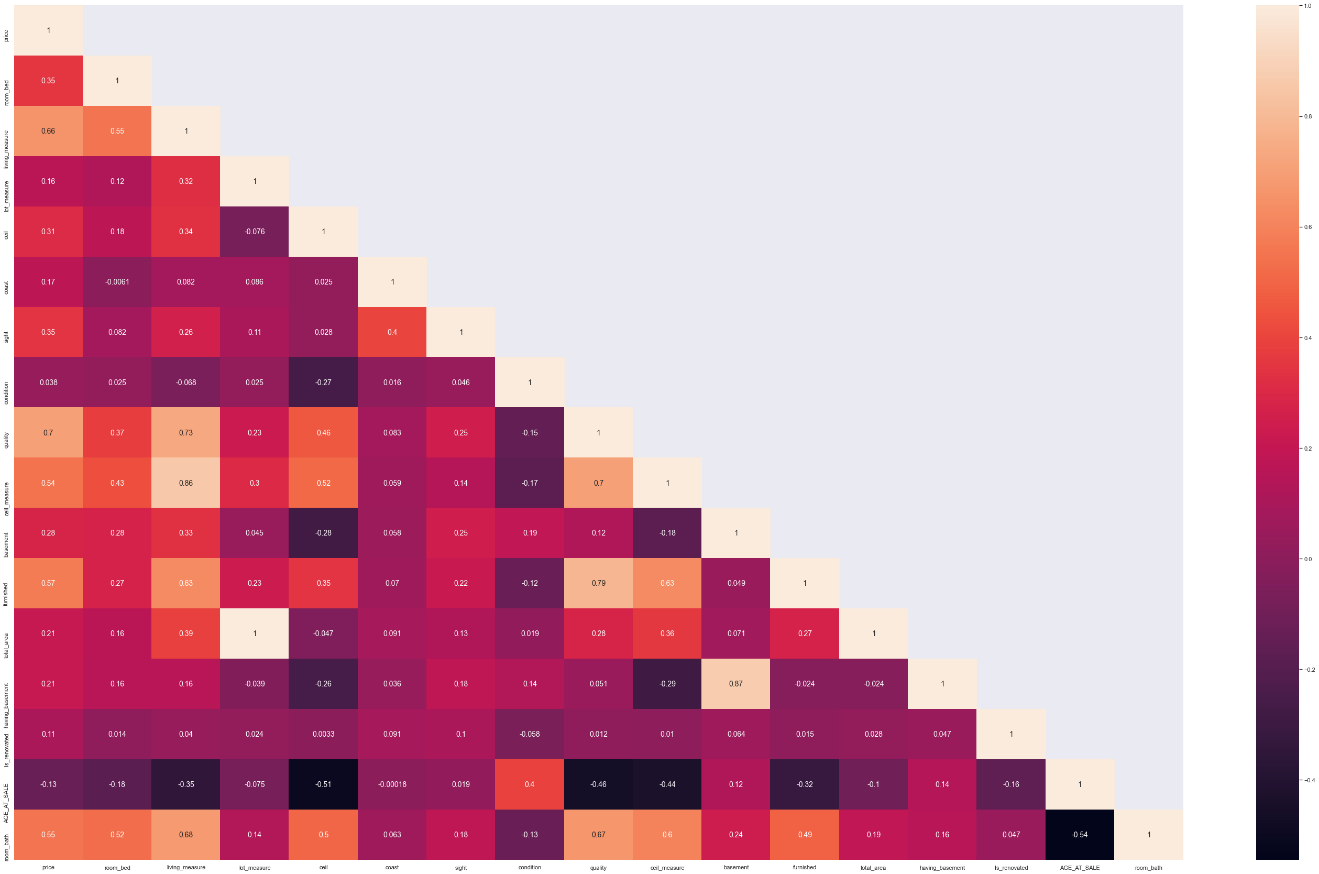
**DATA ANALYSIS AND INTERPRETATION**

**DATA ANALYSIS AND INTERPRETATION**

**3.1 Linear Regression Analysis**

The initial approach utilized linear regression, which excels at modeling linear relationships between features and the target variable (house prices). the model's performance will be evaluated using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. These metrics will provide insights into the model's accuracy and ability to fit the data.

While the linear regression model achieved a baseline performance (R-squared of 0.47), there's room for improvement. This suggests that a linear relationship might not fully capture the complexities influencing house prices.



**3.2 Other Techniques**

Other regularized regression techniques like Ridge and Lasso regression, forest tree decision, Support vector regression (SVR).etc would be used to check for any potential correlated features within the data that can help reduce model complexity and potentially improve generalizability.

The evaluation process highlighted that while linear regression provides a baseline, it might not capture the full picture hence further exploration is required.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**References**

<https://dergipark.org.tr/tr/download/article-file/2042774>

<https://www.researchgate.net/publication/347369616_A_Comparative_Study_on_House_Price_Prediction>

<https://arxiv.org/abs/2204.09050>

<https://medium.com/@blogsupport/decisive-housing-price-prediction-88c8f90a1733>

<https://www.diva-portal.org/smash/get/diva2:1456610/FULLTEXT01.pdf>