

MCA Semester – IV

Research Project – Interim Report

|  |  |
| --- | --- |
| **Name** |  |
| **Project** | HOUSE PRICE PREDICTION |
| **Group** | GROUP 7 |
| **Date of Submission** |  |



**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 Computer Application**

*Submitted by:*

**Student Name**

USN:

(Write your number)

*Under the guidance of:*

Mention your Guide’s Name

(Faculty-JAIN Online)

Jain Online (Deemed-to-be University)

Bangalore

**2022-23**

**DECLARATION**

I, *(Student Name),* hereby declare that the Research Project Report titled *“HOUSE PRICE PREDICTION” has been* prepared by me under the guidance of the *Faculty name.* I declare that this Project work is towards the partial fulfillment of the University Regulations for the award of the degree of Master of Computer Application 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: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: *Name of the Student*

*USN:*

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **Title** | **Page Nos.** |
| Executive Summary | i |
| List of Tables | ii |
| List of Graphs | iii |
| Chapter 1: Introduction and Background | 1-3 |
| Chapter 2: Research Methodology | 4-9 |
| Chapter 3: Data Analysis and Interpretation | 10-15 |
| References |  |
| Annexures |  |
|  |  |
|  |  |

**EXECUTIVE SUMMARY**

Project Title: House Price Prediction

Project Overview:

The House Price Prediction project represents a substantial endeavor aimed at leveraging advanced machine learning techniques to predict house prices accurately. This project is motivated by the ever-evolving real estate market, where informed pricing decisions are crucial for buyers, sellers, and real estate professionals. By developing a reliable prediction model, we seek to empower stakeholders with a valuable tool to navigate the complexities of property pricing.

Key Findings or Highlights:

Initial Data Exploration: The initial data exploration phase has revealed essential insights into the distribution of property features and potential outliers. This phase laid the foundation for subsequent data preprocessing and model development activities.

Data Preprocessing Progress: Fousing on improving data quality and consistency. Key tasks include handling missing values, encoding categorical variables, and handling outliers.

Project Milestones and Achievements:

Data Quality Assurance: The project encountered challenges related to data quality, such as missing values and inconsistencies across data sources. These challenges have been diligently addressed through a combination of data cleaning and imputation strategies.

Data Preprocessing Completion: The data preprocessing phase will be finalized to ensure the high-quality input data necessary for effective model training.

Model Development: Model development , involving experimentation with various machine learning algorithms. Prominent algorithms being explored include Linear Regression, Random Forest,Decision Tree and Gradient Boosting. The goal is to identify the most suitable model for accurate price predictions.

Hyperparameter Tuning: To optimize model performance, hyperparameter tuning is necessary. Techniques such as grid search and cross-validation are being employed to fine-tune model parameters.

Model Evaluation: Rigorous model evaluation is done, utilizing a variety of metrics, including Root Mean Squared Error (RMSE) and R-squared. This evaluation phase ensures the model's accuracy and its ability to deal with new data.

Comprehensive Testing: Extensive cross-validation and testing will be conducted to verify model robustness and reliability in diverse scenarios.

Conclusion:

This data science project House price prediction . We will first build a model using sklearn and linear regression using inner-city home prices datasets . During model building we will cover almost all data science concepts such as data load and cleaning, outlier detection and removal, feature engineering,column transformers ,gridsearch cv for hyperparameter tunning, etc.

Technology and tools wise this project covers,

1 Python

2 Numpy and Pandas for data cleaning

3 Matplotlib and seaborn for data visualization

4 Sklearn for model building

5 Jupyter notebook, visual studio code and pycharm as IDE

|  |  |  |
| --- | --- | --- |
| **List of Tables** | | |
| **Table No.** | **Table Title** | **Page No.** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| **List of Graphs** | | |
| **Graph No.** | **Graph Title** | **Page No.** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**CHAPTER 1**

**INTRODUCTION AND BACKGROUND**

* 1. **Introduction and Background :**

The House Price Prediction project is a data-driven initiative that leverages machine learning to forecast real estate property prices accurately. The project is driven by the recognition of the critical role that housing plays in people's lives, and the profound impact that pricing decisions have on buyers, sellers, and the broader real estate market. By developing a robust prediction model, we aim to empower individuals and professionals with the knowledge needed to make well-informed property pricing decisions.

The dynamics of the real estate market are characterized by a myriad of factors, including location, property size, number of bedrooms, number of baths, and economic conditions. As such, determining the fair market value of a property has often been challenging and subjective. Buyers seek transparency and affordability, while sellers aspire to maximize returns on their investments.

Traditional pricing methods often rely on historical data and expert opinions, but they can be influenced by biases and fail to capture the complexity of contemporary markets. This project addresses these challenges by harnessing the power of machine learning and data analytics to create a predictive model capable of accounting for diverse and dynamic property variables.

* 1. **Problem Statement:**

A house value is simply more than location and square footage. Like the features that make up a  person, an educated party would want to know all aspects that give a house its value. For  example, you want to sell a house and you don’t know the price which you may expect — it can’t  be too low or too high. To find house price you usually try to find similar properties in your  neighborhood and based on gathered data you will try to assess your house price.

* 1. **Objective of Study:**

The objective of the House Price Prediction project is to develop a machine learning model that accurately predicts house prices based on various property features. This project aims to empower people with a reliable tool for making well-informed pricing decisions in the dynamic real estate market ,by taking advantage of all of the feature variables available below,

1. **cid:** Notation for a house. Will not of our use. So we will drop this column
2. **dayhours:** Represents Date, when house was sold.
3. **price:** It's our TARGET feature, that we have to predict based on other featues
4. **room\_bed:** Represents number of bedrooms in a house
5. **room\_bath:** Represents number of bathrooms
6. **living\_measure:** Represents square footage of house
7. **lot\_measure:** Represents square footage of lot
8. **ceil:** Represents number of floors in house
9. **coast:** Represents whether house has waterfront view. It seems to be a categorical variable. We will see in our further data analysis
10. **sight:** Represents how many times sight has been viewed.
11. **condition:** Represents the overall condition of the house. It's kind of rating given to the house.
12. **quality:** Represents grade given to the house based on grading system
13. **ceil\_measure:** Represents square footage of house apart from basement
14. **basement:** Represents square footage of basement
15. **yr\_built:** Represents the year when house was built
16. **yr\_renovated:** Represents the year when house was last renovated
17. **zipcode:** Represents zipcode as name implies
18. **State:** Represents state
19. **City**:Represents City
20. **lat:** Represents Lattitude co-ordniates
21. **long:** Represents Longitude co-ordinates
22. **living\_measure15:** Represents square footage of house, when measured in 2015 year as house area may or may not changed after renovation if any happened
23. **lot\_measure15:** Represents square footage of lot, when measured in 2015 year as lot area may or may not change after renovation if any done
24. **furnished:** Tells whether house is furnished or not. It seems to be categorical variable as description implies
25. **total\_area:** Represents total area i.e. area of both living and lot
    1. **Company and industry overview:**

The House Price Prediction project operates at the intersection of the real estate and technology industries, which are both undergoing transformative changes in recent years.The real estate industry is a cornerstone of the global economy, encompassing residential, commercial, and industrial properties. It plays a pivotal role in wealth creation, investment, and housing provision. However, it is a sector characterized by complexity, local variations, and the need for timely and accurate pricing information.

The House Price Prediction project is a flagship initiative within Company portfolio. This project exemplifies our dedication to leveraging advanced data science techniques to address real-world challenges and deliver practical solutions that benefit a wide range of stakeholders.

* 1. **Overview of Theoretical Concepts:**

1. Feature Engineering:

Feature engineering involves the process of selecting, creating, or transforming features (variables) in the dataset to improve the performance of a predictive model.

Techniques may include one-hot encoding for categorical variables, scaling for numerical features, and generating new features based on domain knowledge.

2. Data Preprocessing:

Data preprocessing is a crucial step that includes cleaning and transforming raw data into a suitable format for modeling.

Tasks may involve handling missing values, dealing with outliers, and ensuring data consistency.

3. Cross-Validation:

Cross-validation is a technique used to assess a model's performance and generalization ability. It involves splitting the dataset into multiple subsets (folds), training the model on some folds and evaluating it on others.

4. Hyperparameter Tuning:

Hyperparameter tuning involves finding the best set of hyperparameters for a machine learning model to optimize its performance.

Techniques such as grid search and random search are used to systematically explore different hyperparameter combinations.

5. Evaluation Metrics:

Evaluation metrics are used to quantify how well a predictive model performs. Common metrics for regression tasks include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared.

**CHAPTER 2**

**EXPLORATORY DATA ANALYSIS**

**2.1 Scope of the Study**

**2.2 Methodology**

**2.2.1 Research Design**

**2.2.2 Data Collection**

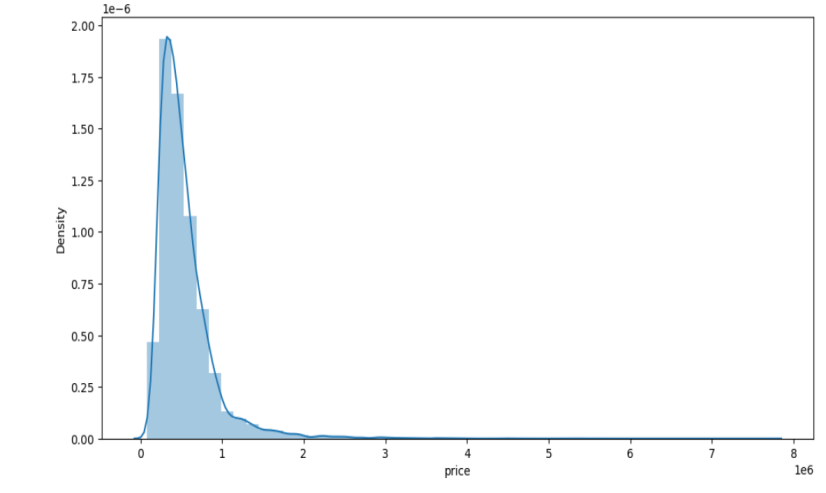
**2.2.3 Sampling Method (if applicable)**

**2.2.4 Data Analysis Tools**

**2.3 Period of Study**

**2.4 Utility of Research**

**2.1 Univariate Analysis:**

****

**2.1 Scope of the Study**

**2.2 Methodology**

**2.2.1 Research Design**

**2.2.2 Data Collection**

**2.2.3 Sampling Method (if applicable)**

**2.2.4 Data Analysis Tools**

**2.3 Period of Study**

**2.4 Utility of Research**

**(Note: This must include:**

**Summary of the Approach to EDA and Pre-processing- Include any insightful visualization you have teased out of the data. If you’ve identified particularly meaningful features, interactions or summary data, share them and explain what you noticed. Visual displays are powerful when used well, so think carefully about what information the display conveys.)**

**CHAPTER 3**

**DATA ANALYSIS AND INTERPRETATION**

**DATA ANALYSIS AND INTERPRETATION**

**(Note: What are the approaches you can take to improve your model? Can you do some feature selection, data manipulation and model improvements?**

**Provide your code and as much as visualizations you can share to describe what you have done so far.)**

**ANNEXURE (if any)**

**The questionnaires, financial statements and any other relevant document can be put here. The annexures have to be numbered in case there are more than one annexure.**