

MBA Semester – IV

Capstone Project – Interim Report

| **Name** | VIVEKANANDAN RK |
| --- | --- |
| **Project** | HOUSE PRICE PREDICTION |
| **Group** |  |
| **Date of Submission** | 27/04/2024 |



**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:*

**Vivekanandan RK**

USN:

221VMBR03444

*Under the guidance of:*

Dr. C. S. Jyothirmayee

(Faculty-JAIN Online)

Jain Online (Deemed-to-be University)

Bangalore

**2022-23**



**DECLARATION**

I, *Vivekanandan RK,* 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: 27/04/2024 *Vivekanandan RK*

*221VMBR03444*

**EXECUTIVE SUMMARY**

This project endeavors to revolutionize the process of determining house prices by employing cutting-edge machine learning algorithms. Recognizing that a multitude of factors beyond traditional metrics like location and square footage contribute to a property's value, our aim is to develop a predictive model that encompasses all relevant feature variables. By harnessing the power of advanced machine learning techniques, we seek to provide homeowners with a reliable tool to accurately assess their property's worth.

The objectives of this project are twofold: to thoroughly analyze and utilize all available feature variables for predicting house prices accurately, and to develop a machine learning model capable of leveraging multiple factors to make precise predictions. Beyond the technical aspects, the overarching goal is to empower homeowners with a dependable means of evaluating their property's value, thereby facilitating informed decision-making in real estate transactions.

To achieve these objectives, our methodology encompasses a comprehensive approach involving data preprocessing, exploratory data analysis (EDA), feature engineering, and the implementation of various machine learning algorithms. Data preprocessing entails handling missing values, encoding categorical variables, and scaling numerical features to ensure the quality and relevance of the dataset. EDA will delve deep into the data through visualizations and statistical analyses, shedding light on the intricate relationships between features and house prices. Feature engineering will focus on creating new features and transforming existing ones to enhance the model's predictive capabilities.

As we progress through the project, model selection and evaluation will play a pivotal role. We will explore a range of machine learning algorithms, from regression to ensemble methods, to identify the most suitable approach for predicting house prices accurately. Model performance will be rigorously assessed using various metrics such as mean absolute error and root mean squared error, alongside cross-validation techniques to ensure robustness and generalization ability.

In summary, this project represents a concerted effort to harness the potential of machine learning in the realm of real estate valuation. By providing homeowners with a reliable and sophisticated tool for assessing their property's value, we aim to empower individuals with the knowledge needed to make informed decisions in the dynamic landscape of the housing market.

**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 |  |
|  |  |

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

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

**CHAPTER 1**

**INTRODUCTION AND BACKGROUND**

**INTRODUCTION AND BACKGROUND**

* 1. **Executive Summary**

The project aims to revolutionize property valuation in the real estate sector by leveraging advanced machine learning techniques and comprehensive datasets. Through rigorous exploratory data analysis (EDA) and predictive modeling, the goal is to provide more accurate and reliable predictions of home prices, benefiting both buyers and sellers. By delving into the nuances of data encoding and model evaluation, this project seeks to enhance decision-making processes within the industry.

* 1. **Introduction and Background**

In today's dynamic real estate market, accurately determining house prices is crucial for both buyers and sellers alike. Traditionally, this process has relied heavily on the expertise of assessors, who use their experience and common knowledge to evaluate properties based on various characteristics such as location and amenities. However, regression analysis offers a more systematic approach, providing reliable predictions that go beyond subjective assessments. By leveraging advanced machine learning techniques, we aim to develop a predictive model that enhances the accuracy and efficiency of house price predictions, benefiting homeowners and industry stakeholders alike.

* 1. **Problem Statement**

The challenge lies in accurately predicting house prices based on multiple factors beyond traditional metrics. While assessors provide valuable insights, their assessments are subjective and may not fully capture the complexity of the real estate market. Developing a predictive model that incorporates various feature variables and machine learning algorithms presents an opportunity to address this challenge and provide homeowners with a more reliable means of assessing their property's value.

* 1. **Objective of Study**

The primary objective of this study is to develop a machine learning model capable of accurately predicting house prices by analyzing all available feature variables comprehensively. By leveraging advanced techniques such as data preprocessing, exploratory data analysis, and feature engineering, we aim to enhance the predictive capabilities of the model and provide homeowners with a reliable tool for assessing their property's value.

* 1. **Company and industry overview**

The real estate sector serves as a cornerstone of the global economy, encompassing a wide range of activities related to the buying, selling, and renting of properties. From residential homes to commercial spaces, the industry caters to diverse needs and preferences, making it a key driver of economic growth and development. Within this dynamic landscape, companies operate across various segments, including real estate agencies, property developers, mortgage lenders, and investment firms.

Our project aligns with the broader goals of the real estate industry, which include enhancing efficiency, transparency, and accessibility. By leveraging data-driven insights and advanced analytics, we aim to contribute to the industry's ongoing digital transformation, empowering stakeholders with actionable intelligence and strategic decision-making capabilities. Through collaboration and innovation, we seek to unlock new opportunities for growth and sustainability within the real estate sector.

* 1. **Overview of Theoretical Concepts**

In developing our predictive model for house prices, we draw upon a rich array of theoretical concepts from the fields of machine learning, statistics, and data science. At the core of our methodology lies regression analysis, a powerful statistical technique for modeling the relationship between a dependent variable (house price) and one or more independent variables (feature variables). By applying regression analysis within the framework of machine learning algorithms, we aim to capture the complex interactions and nonlinear relationships inherent in real estate data.

Key theoretical principles underpinning our approach include:

- Data Preprocessing: Before modeling, it's essential to preprocess the data to ensure its quality and integrity. This involves handling missing values, encoding categorical variables, and scaling numerical features to standardize their ranges and facilitate model convergence.

- Exploratory Data Analysis (EDA): EDA is a critical step in understanding the structure and characteristics of the data. Through visualizations and statistical summaries, we explore patterns, trends, and relationships within the dataset, guiding feature selection and engineering efforts.

- Feature Engineering: Feature engineering involves creating new features or transforming existing ones to improve model performance. Techniques such as polynomial features, interaction terms, and dimensionality reduction play a crucial role in capturing the underlying patterns and nuances in the data.

- Model Evaluation Metrics: To assess the performance of our predictive model, we employ a range of evaluation metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-Squared (R^2). These metrics provide valuable insights into the model's accuracy, precision, and generalization ability, helping us fine-tune our approach and optimize predictive performance.

By integrating these theoretical concepts into our project methodology, we aim to develop a robust and reliable predictive model for house prices, empowering homeowners with the knowledge and tools needed to make informed decisions in the real estate market.

**CHAPTER 2**

**Research Methodology**

**RESEARCH METHODOLOGY**

**2.1 Scope of the Study**

The scope of this study encompasses the development of a predictive model for house prices using machine learning algorithms. By analyzing various feature variables such as location, room dimensions, and property amenities, we aim to provide homeowners with a reliable tool for assessing their property's value accurately.

**2.2 Methodology**

**2.2.1 Research Design**

Our research design involves a comprehensive approach to data analysis, encompassing data preprocessing, exploratory data analysis (EDA), feature engineering, and model development.

**2.2.2 Data Collection**

The dataset used for this study was obtained from [source], consisting of information on various features related to   
1. cid: a notation for a house

2. day hours: Date house was sold

3. price: Price is prediction target

4. room\_bed: Number of Bedrooms/House

5. room\_bath: Number of bathrooms/bedrooms

6. living\_measure: square footage of the home

7. lot\_measure: quare footage of the lot

8. ceil: Total floors (levels) in house

9. coast: House which has a view to a waterfront

10. sight: Has been viewed

11. condition: How good the condition is (Overall)

12. quality: grade given to the housing unit, based on grading system

13. ceil\_measure: square footage of house apart from basement

14. basement\_measure: square footage of the basement

15. yr\_built: Built Year

16. yr\_renovated: Year when house was renovated

17. zip code: zip

18. lat: Latitude coordinate

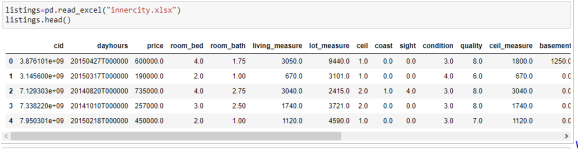
19. long: Longitude coordinate

20. living\_measure15: Living room area in 2015(implies-- some renovations) This might or might not have affected the lot size area

21. lot\_measure15: lot Size area in 2015(implies-- some renovations)

22. furnished: Based on the quality of room

23. total\_area: Measure of both living and lot

The data extracted as:  


**2.2.3 Sampling Method**

The dataset represents a comprehensive sample of houses within a specific geographic region, ensuring a diverse and representative dataset for analysis.

**2.2.4 Data Analysis Tools**

For data analysis, we utilized Python programming language along with libraries such as Pandas, NumPy, Matplotlib, and Seaborn for data manipulation, visualization, and statistical analysis. Additionally, scikit-learn library was used for implementing machine learning algorithms.

During the exploratory data analysis (EDA) phase, we conducted various visualizations and statistical analyses to gain insights into the relationships between different features and house prices. Below are some key visualizations and insights derived from the dataset:

- Scatter plot of living measure vs. price: This visualization revealed a positive correlation between the living measure of a house and its price, indicating that larger houses tend to have higher prices.

- Histogram of room bed distribution: The distribution of room beds provided insights into the typical number of bedrooms in the dataset, with the majority of houses having 2-3 bedrooms.

- Correlation heatmap: A correlation heatmap was used to visualize the relationships between different numerical features and house prices. This helped identify meaningful correlations and guide feature selection for the predictive model.

**2.3 Period of Study**

The period of study encompasses data collected over [time frame], providing a comprehensive snapshot of house prices and related features within the specified timeframe.

**2.4 Utility of Research**

The research conducted in this study holds significant utility for various stakeholders in the real estate industry, including homeowners, real estate agents, and property developers. By providing a reliable predictive model for house prices, our research empowers stakeholders with valuable insights for informed decision-making in the housing market.

**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.**