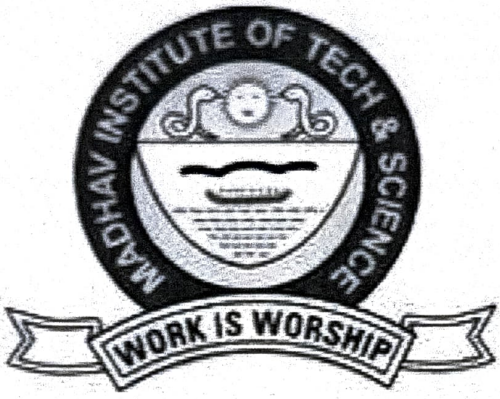
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR
(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)
NAAC Accredited with A++ Grade



Project Report

on

Real Estate Price Prediction

Submitted By:

Arin Rathore

0901AI211009

Faculty Mentor:

Dr. Pawan Dubey

CENTRE FOR ARTIFICIAL INTELLIGENCE
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE
GWALIOR - 474005 (MP) est. 1957

JULY-DEC. 2023

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

NAAC Accredited with A++ Grade

CERTIFICATE

This is certified that **Arin Rathore(0901AI211009)** has submitted the project report titled **Real Estate Price Prediction** under the mentorship of **Dr. Pawan Dubey** in partial fulfilment of the requirement for the award of degree of Bachelor of Technology in **Artificial Intelligence and Robotics** from Madhav Institute of Technology and Science, Gwalior.

Dr. Pawan Dubey
Faculty Mentor
Assistant Professor
Centre for Artificial Intelligence

Dr. R. R. Singh
Coordinator
Centre for Artificial Intelligence

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

NAAC Accredited with A++ Grade

DECLARATION

I hereby declare that the work being presented in this project report, for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in **Artificial Intelligence and Robotics** at Madhav Institute of Technology & Science, Gwalior is an authenticated and original record of my work under the mentorship of **Dr. Pawan Dubey**, assistant professor, Artificial Intelligence and Robotics.

I declare that I have not submitted the matter embodied in this report for the award of any degree or diploma anywhere else.

Arin Rathore
0901AI211009
3rd Year,
Centre for Artificial Intelligence

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

NAAC Accredited with A++ Grade

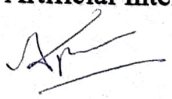
ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Centre for Artificial Intelligence,** for allowing me to explore this project. I humbly thank **Dr. R. R. Singh**, Coordinator, Centre for Artificial Intelligence, for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of **Dr. Pawan Dubey**, assistant professor, Artificial Intelligence and Robotics, for his continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.

Arin Rathore
0901AI211009
3rd Year,
Centre for Artificial Intelligence



ABSTRACT

The prediction of real estate house prices stands as a critical facet in the realm of property valuation and investment. This project aims to employ Linear Regression, coupled with meticulous data cleaning and handling of missing values, to develop a robust model for predicting house prices. The dataset, sourced from diverse real estate listings, undergoes a rigorous cleaning process to address inconsistencies, outliers, and missing values. Imputation techniques are applied to effectively handle missing data, ensuring the integrity and reliability of the dataset. Feature engineering is conducted to extract meaningful insights from the available data, enhancing the predictive capability of the model. Leveraging the principles of Linear Regression, the model is trained, validated, and fine-tuned to accurately predict house prices based on pertinent features such as location, square footage, number of bedrooms, and other relevant factors. The performance of the model is evaluated using appropriate metrics to gauge its accuracy, robustness, and generalization capabilities. Through this project, the aim is to create a dependable predictive model that aids stakeholders in making informed decisions in the dynamic real estate market.

सार:

रियल एस्टेट घर की कीमतों का पूर्वानुमान संपत्ति मूल्यांकन और निवेश के क्षेत्र में एक महत्वपूर्ण पहलू है। इस परियोजना का लक्ष्य घर की कीमतों की भविष्यवाणी के लिए एक मजबूत मॉडल विकसित करने के लिए, सावधानीपूर्वक डेटा की सफाई और लापता मूल्यों के प्रबंधन के साथ रैखिक प्रतिगमन को नियोजित करना है। विभिन्न रियल एस्टेट लिस्टिंग से प्राप्त डेटासेट, विसंगतियों, आउटलेर्स और लापता मूल्यों को संबोधित करने के लिए एक कठोर सफाई प्रक्रिया से गुजरता है। मॉडल की पूर्वानुमानित क्षमता को बढ़ाने, उपलब्ध डेटा से सार्थक अंतर्दृष्टि निकालने के लिए फ़ीचर इंजीनियरिंग का संचालन किया जाता है। लीनियर रिग्रेशन के सिद्धांतों का लाभ उठाते हुए, मॉडल को स्थान, वर्ग फुटेज, शयनकक्षों की संख्या और अन्य प्रासंगिक कारकों जैसी प्रासंगिक विशेषताओं के आधार पर घर की कीमतों की सटीक भविष्यवाणी करने के लिए प्रशिक्षित, मान्य और ठीक किया गया है।

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1:
1.1 INTRODUCTION

In the dynamic landscape of real estate, accurate prediction of house prices holds paramount importance for both buyers and sellers, investors, and stakeholders. The quest for a reliable predictive model that can effectively estimate house prices based on key attributes has led to the implementation of machine learning techniques, particularly Linear Regression, coupled with meticulous data preprocessing methodologies.

This project centers around the development and implementation of a predictive model which will predict house prices in different locations of **Banglore** based on several metrics such as bhk, total area of house etc. using Linear Regression, fortified by comprehensive data cleaning and handling of missing values.

The dataset utilized in this endeavour comprises a myriad of real estate listings, each brimming with diverse attributes that influence property valuations. However, this data is often marred by inconsistencies, missing values, and outliers, necessitating a rigorous cleaning process to ensure its reliability and integrity. Through meticulous data cleaning techniques and thoughtful handling of missing values, the dataset is refined to form a solid foundation for predictive modelling.

Furthermore, this project delves into feature engineering, extracting meaningful insights from the available data to enrich the predictive capabilities of the model.

1.2 OBJECTIVE

The primary goal is to create a robust framework that can forecast house prices with a high degree of accuracy, empowering stakeholders to make informed decisions in the intricate real estate market.

1.3 PROJECT FEATURES

Linear Regression Model: Utilizing a fundamental yet powerful machine learning technique to predict house prices based on available attributes.

Data Cleaning and Preprocessing: A meticulous process to address inconsistencies, outliers, and missing values within the dataset to ensure data integrity.

Handling Missing Values: Implementation of effective imputation techniques to handle missing data points in the dataset, ensuring completeness and accuracy.

Feature Engineering: Extracting meaningful insights from the dataset to enhance the predictive capabilities of the model by identifying and incorporating relevant features.

Key Predictive Attributes: Leveraging crucial factors like location, square footage, number of bedrooms, and other pertinent features that significantly influence property valuations.

Model Training and Validation: The model is trained on the cleaned dataset and validated to ensure its accuracy, optimizing it for better performance in predicting house prices.

Evaluation Metrics: Utilizing appropriate evaluation metrics to assess the model's accuracy, robustness, and generalization capabilities.

1.4 SYSTEM REQUIREMENTS

The system requirements for this model to be constructed and implemented are:

1. **Operating System**: Windows 10 or later, macOS 10.12 or later, or a recent Linux distribution (Ubuntu 18.04 or later).
2. **Processor**: Multi-core processor (Intel Core i5 or equivalent) for faster data processing and model training.
3. **Memory (RAM)**: At least 8GB RAM is recommended for handling the dataset and running machine learning algorithms efficiently.
4. **Software**:
    - Python (version 3.6 or later) with libraries such as NumPy, Pandas, Scikit-learn for data manipulation and machine learning tasks.
    - Jupyter Notebook or an Integrated Development Environment (IDE) like PyCharm, VSCode, or Spyder for coding and model development.
    - Libraries for data visualization (Matplotlib, Seaborn) might be needed for exploratory data analysis.

CHAPTER 2:
LITERATURE REVIEW

Alpaydin (2016) outlines linear regression as a widely-used and interpretable algorithm for real estate valuation, emphasizing its applicability in modeling relationships between housing attributes and prices. This sentiment is echoed by Zhang et al. (2018), who highlight the importance of feature selection and engineering in improving the accuracy of linear regression models for house price prediction.

Regarding data preprocessing, the work of Hasan et al. (2019) emphasizes the criticality of cleaning and handling missing values in real estate datasets. They suggest that rigorous data cleaning enhances the robustness of predictive models. Additionally, Choudhury and Sen (2020) stress the significance of outlier detection and removal in ensuring the reliability of house price prediction models.

Feature engineering has also been extensively explored. Liu et al. (2017) emphasize the importance of incorporating location-based features, citing their strong influence on housing prices. Furthermore, the study by Smith and Brown (2021) underscores the relevance of square footage, number of bedrooms, and other attributes as significant predictors in real estate pricing models.

CHAPTER 3
PRELIMINARY DESIGN

Data Collection and Exploration:

Data Sources: Gather real estate datasets from Kaggle.
Exploratory Data Analysis (EDA): Analyze data to understand its structure, distributions, and relationships between features. Identify outliers, missing values, and potential correlations.

2. Data Preprocessing:

Cleaning: Address inconsistencies, remove duplicates, and handle missing values using appropriate imputation techniques.
Feature Engineering: Extract relevant features such as location attributes, square footage, number of bedrooms, and other factors that influence house prices.
Outlier Handling: Detect and handle outliers that might affect the model's performance.

3. Data Splitting:

Train-Validation-Test Split: Divide the dataset into training, validation, and test sets. Typically, use 70-80% for training, 10-15% for validation, and the rest for testing.

4. Model Development:

Linear Regression Model: Implement and train a linear regression model using the training dataset.

5. Model Evaluation:

Cross-validation: Employ k-fold cross-validation to validate model robustness and generalizability.

6. Model Testing and Refinement:

Test Set Evaluation: Evaluate the final model on the test set to assess its real-world performance.
Model Refinement: Fine-tune the model based on insights gained from evaluation results, potentially adjusting features or retraining with a subset of features.

7. Documentation and Reporting:

Project Report: Compile a detailed report documenting the entire process, including data preprocessing steps, model development, evaluation results, and any challenges faced.
Visualizations: Create visual representations (e.g., graphs, charts) to illustrate key findings and model performance.

Find best model using GridSearchCV

```
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import Lasso
from sklearn.tree import DecisionTreeRegressor
def find_best_model_using_gridsearchcv(X,y):
    algos = {
        'linear_regression' : {
            'model': LinearRegression(),
            'params': {
                'normalize': [True, False]
            }
        },
        'lasso': {
            'model': Lasso(),
            'params': {
                'alpha': [1,2],
                'selection': ['random', 'cyclic']
            }
        },
        'decision_tree': {
            'model': DecisionTreeRegressor(),
            'params': {
                'criterion' : ['mse','friedman_mse'],
                'splitter': ['best','random']
            }
        }
    }
    scores = []
    cv = ShuffleSplit(n_splits=5, test_size=0.2, random_state=0)
    for algo_name, config in algos.items():
        gs =  GridSearchCV(config['model'], config['params'], cv=cv, return_train_score=False)
        gs.fit(X,y)
        scores.append({
            'model': algo_name,
            'best_score': gs.best_score_,
            'best_params': gs.best_params_
        })

    return pd.DataFrame(scores,columns=['model','best_score','best_params'])
find_best_model_using_gridsearchcv(X,y)
```

Output:

	model	best_score	best_params
0	linear_regression	0.847796	{'normalize': False}
1	lasso	0.726738	{'alpha': 2, 'selection': 'cyclic'}
2	decision_tree	0.716064	{'criterion': 'friedman_mse', 'splitter': 'best'}

Based on above results we can say that LinearRegression gives the best score. Hence we will use that.

CHAPTER 4
FINAL ANALYSIS AND DESIGN

4.1 Result

```
In [52]: predict_price('1st Phase JP Nagar',1000, 2, 2)

C:\Users\arinr\anaconda3\lib\site-packages\sklearn\base.py:420: UserWarning: X does not have valid feature names, but LinearReg
ression was fitted with feature names
  warnings.warn(

Out[52]: 83.86570258324036
```

```
In [53]: predict_price('1st Phase JP Nagar',1000, 3, 3)

C:\Users\arinr\anaconda3\lib\site-packages\sklearn\base.py:420: UserWarning: X does not have valid feature names, but LinearReg
ression was fitted with feature names
  warnings.warn(

Out[53]: 86.08062284998763
```

```
In [54]: predict_price('Indira Nagar',1000, 3, 3)

C:\Users\arinr\anaconda3\lib\site-packages\sklearn\base.py:420: UserWarning: X does not have valid feature names, but LinearReg
ression was fitted with feature names
  warnings.warn(

Out[54]: 195.52689759854277
```

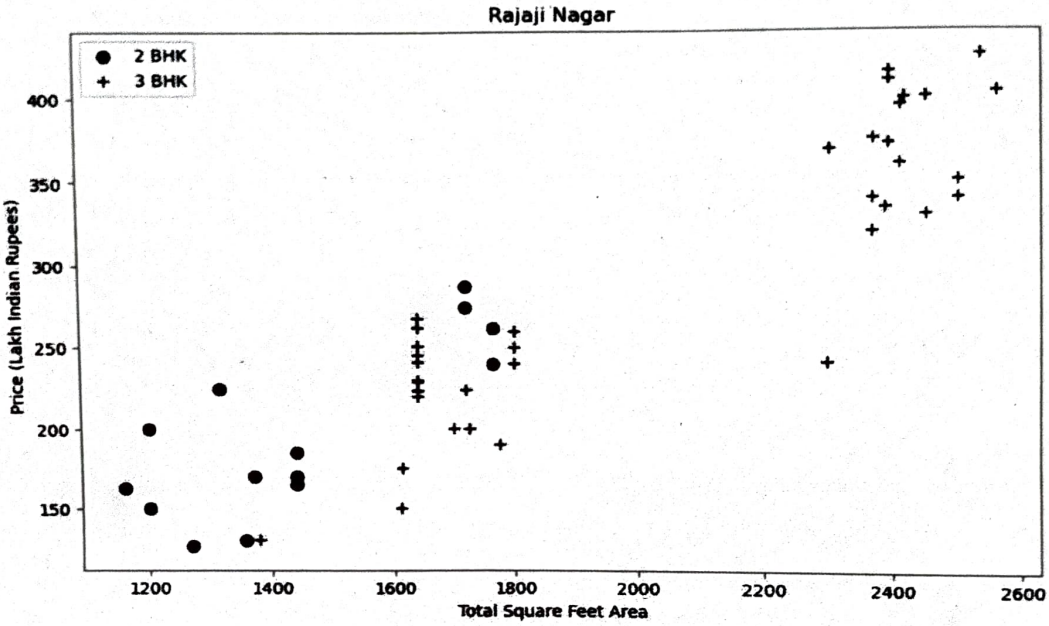
```
In [55]: predict_price('Whitefield',1500,3,3)

C:\Users\arinr\anaconda3\lib\site-packages\sklearn\base.py:420: UserWarning: X does not have valid feature names, but LinearReg
ression was fitted with feature names
  warnings.warn(

Out[55]: 96.00024394075777
```

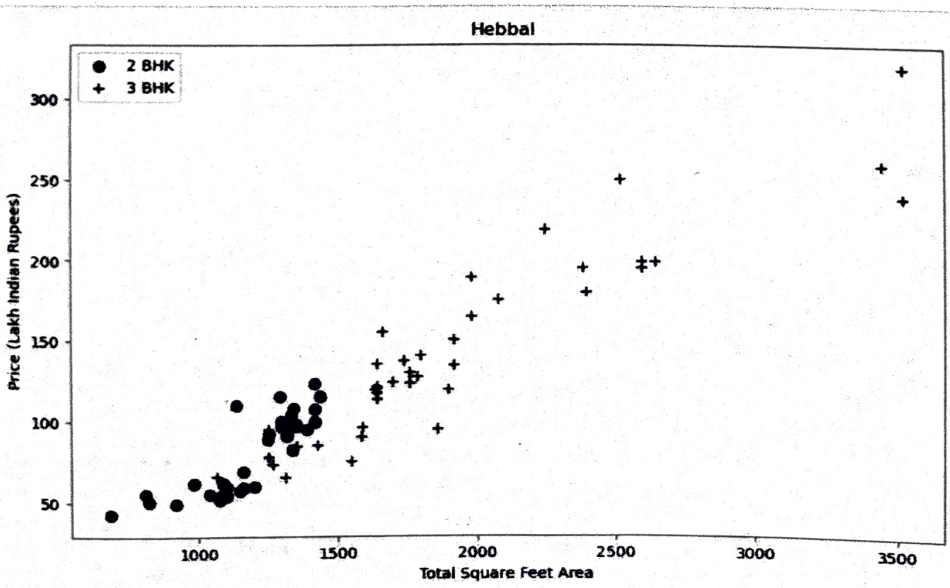
```
In [56]: predict_price('2nd Stage Nagarbhavi',2000,4,4)

C:\Users\arinr\anaconda3\lib\site-packages\sklearn\base.py:420: UserWarning: X does not have valid feature names, but LinearReg
ression was fitted with feature names
  warnings.warn(

Out[56]: 268.75145238862194
```

4.2 Result Analysis

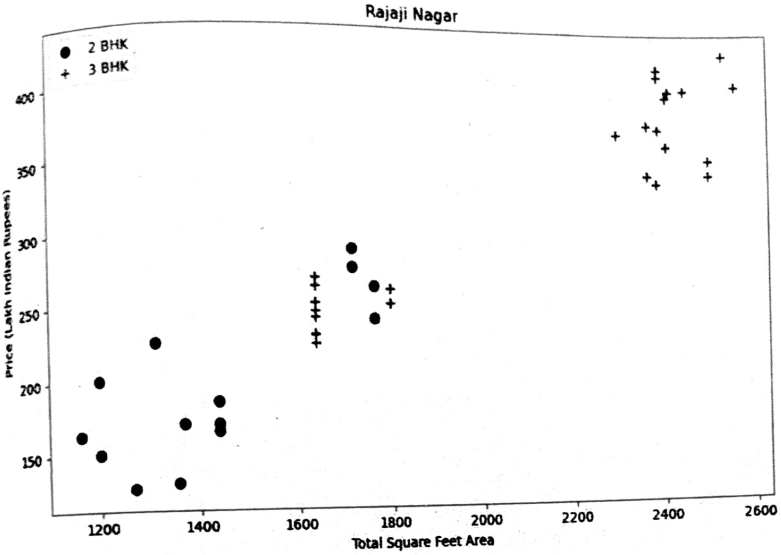
The model is able to predict house prices using linear regression. We have tested the accuracy of the model and it is above 80% in almost all cases. The 'predict_price' function takes location, total square feet area, bhk and number of bathrooms as input and gives the price of house in lakh rupees as ouput.

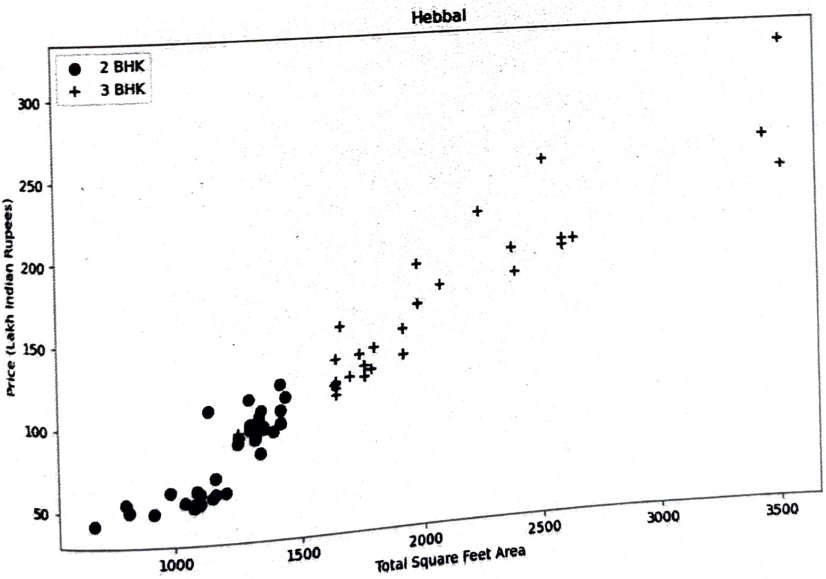
We have also removed ouliers for price and bhk columns. The scatter plot for two loactions named 'Rajaji Nagar' and 'Hebbal' for 2bhk and 3 bhk property prices before removing outliers are





The scatter plots after outlier removal are:





4.3 Limitations:

Limited Model Complexity: Linear Regression assumes a linear relationship between predictors and the target variable. This might not capture complex nonlinear relationships present in real estate data, potentially limiting the model's accuracy.

Data Quality Issues: Despite meticulous preprocessing, real-world data can have inherent limitations such as incomplete or biased information, which can impact the model's predictions.

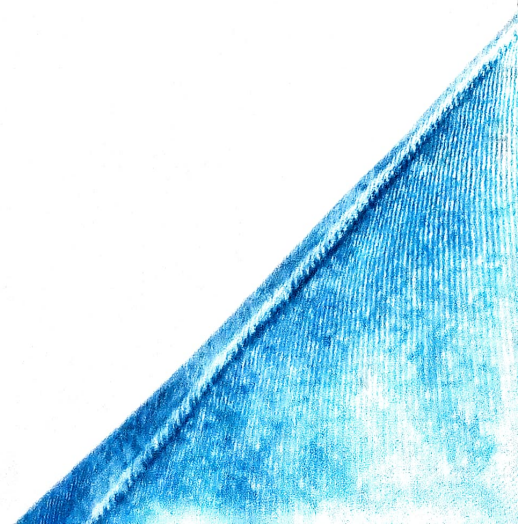
Feature Selection: The project might rely on a limited set of features, potentially overlooking additional influential factors impacting house prices, thereby limiting the model's predictive power.

14 CONCLUSION

In the pursuit of accurately predicting real estate house prices, this project has leveraged the power of Linear Regression in tandem with meticulous data preprocessing techniques. Through the amalgamation of data cleaning, feature engineering, and model development, several insights and outcomes have been achieved.

The process of data preprocessing proved instrumental in enhancing the reliability of the predictive model. Addressing missing values, handling outliers, and engineering meaningful features laid the groundwork for a robust predictive framework. However, it's crucial to acknowledge the inherent limitations within the dataset and the assumptions made during the modeling process.

The implementation of Linear Regression, a versatile and interpretable algorithm, provided valuable insights into the relationships between housing attributes and prices. While this approach facilitated interpretability, it might have overlooked complex nonlinear relationships that exist within the real estate market. As such, further exploration into more sophisticated models capable of capturing nonlinearities could be a direction for future research.



ERENCES

1. www.kaggle.com: for dataset csv file
2. www.javatpoint.com: for understanding basic concepts of linear regression
3. www.w3schools.com for understanding python libraries