

Predicting Pune House Prices: A Machine Learning Approach

Levansh Bhan¹, Vedant Vispute²

¹⁻² Mrs Prachi Karale, Department of Computer Engineering, Savitribai Phule Pune University, Pune, MH, India

Abstract - We propose the implementation of a machine learning model for predicting house prices in Pune, India. This model integrates data science and web development techniques, and we have deployed it on Elastic Beanstalk. Our project aims to address the issue of fluctuating and potentially exaggerated housing prices by focusing on genuine factors. We aim to evaluate the pricing based on essential criteria commonly considered in establishing house prices. The primary objective of this project is to gain hands-on experience in Python, data analytics, machine learning, and artificial intelligence, contributing to the advancement of predictive analytics in the real estate domain.

Key Words: Pune House Price Prediction (PHPP), Pandas, NumPy, Matplotlib, Exploratory Data Analysis (EDA), Data Cleaning, Feature Engineering, Dimensionality Reduction, Data Visualisation, PythonAnywhere, Anaconda, Flask, Scikit-learn.

1. INTRODUCTION

1.1 Context

This research project was undertaken with the aim of satisfying our curiosity and gaining practical experience in the field of machine learning. By immersing ourselves in this project, we sought to enhance our knowledge and skills in developing machine-learning models firsthand.

1.2 Motivation

Our profound interest in the realm of machine learning has led us to embark on this research project, providing us with a valuable opportunity to delve deeper into this subject and rekindle our passion for it. The immense potential of machine learning in generating predictions, and forecasts, and enabling autonomous learning capabilities is awe-inspiring, offering limitless possibilities across various domains. With its wide-ranging applicability in fields such as finance, medicine, and countless others, we deliberately chose to centre our research idea around the fascinating world of machine learning.

1.3 Objective

In our inaugural research endeavour, we aimed to design a project that would provide comprehensive instruction by thoroughly exploring each stage of the machine learning

process, ensuring a deep understanding of its intricacies. To accomplish this, we deliberately selected the Pune Real Estate Prediction task, commonly referred to as a "toy problem." Although not directly aligned with pressing scientific concerns, such problems serve as valuable tools for demonstration and practical application. Our primary objective revolved around developing a robust forecasting model for the price of a specific apartment, incorporating an array of pertinent "features" that would be elucidated in subsequent sections of our research.

2. Literature Survey

Real estate property holds immense significance as it represents not only a person's primary desire but also their wealth and social standing in contemporary society. Investing in real estate is often seen as a lucrative option due to the generally stable nature of property values. Fluctuations in real estate prices impact a wide range of stakeholders, including home investors, bankers, policymakers, and others. Anticipating real estate prices is crucial for economic indicators, considering that the Asian country ranks second globally in terms of the number of households. However, past economic downturns have revealed the unpredictability of real estate costs, which are closely tied to the overall economic conditions of a region. Unfortunately, standardised approaches to accurately forecast real estate property values are lacking.

In our research, we conducted an extensive review of articles and discussions on machine learning applications for housing price prediction. One notable publication focuses on house price prediction utilising machine learning and neural networks, aiming for minimal error and maximum accuracy. Another study titled "Hedonic models based on price data from Belfast" explores the identification of submarkets and residential valuation, shedding light on the evaluation process involving the selection of comparable evidence and the quality of variables that influence property values. The article delves into understanding current trends in house prices and homeownership, emphasising the role of feedback mechanisms and social influences in shaping perceptions of real estate as a crucial market investment.

3. METHODOLOGY

3.1 Data Collection

The data used in this research study was collected from home prices in Pune. The dataset comprises multiple variables, including area type, availability, location, number of bedrooms (BHK), society, total square feet area, number of bathrooms, and number of balconies. These variables provide comprehensive information for the analysis and prediction of house prices in Pune.

3.2 Linear Regression

In the realm of supervised learning, linear regression stands as a fundamental technique. Its primary purpose is to predict the value of a dependent variable (Y) by leveraging an independent variable (X). Linear regression establishes a relationship between the input (X) and the output (Y), serving as one of the most widely recognised and comprehensively studied machine learning algorithms. Notable linear regression models include simple linear regression, ordinary least squares, Gradient Descent, and Regularisation.

3.3 Decision Tree Regression

In the context of predictive modelling, decision tree regression serves as a valuable tool for generating meaningful and continuous output by training a tree-structured model. The fundamental principles underlying decision trees encompass concepts such as Maximising Information Gain, Classification trees, and Regression trees. The core idea revolves around the recursive partitioning process used to construct decision trees. Starting from the root node, which acts as the parent node, each node can be further divided into child nodes, potentially serving as parent nodes for subsequent offspring nodes. The informative features play a crucial role in this process, as they are determined based on maximising information gain. This objective function ensures the optimisation of the tree learning method to achieve accurate predictions and meaningful insights.

3.4 Classification of Trees

Classification trees are employed to predict the categorisation of an object into different classes based on one or multiple predictor variables. These trees provide a framework for effectively assigning categorical labels to the target variable, enabling accurate classification of data instances.

3.5 Regression Trees

Regression trees are versatile models that can handle both continuous and categorical input variables. In the domain of regression problems, various machine learning

algorithms have been explored, and the Decision Tree approach has shown the lowest loss. The Decision Tree model exhibits exceptional performance, as evidenced by an impressive R-Squared value of 0.998. This high value signifies the model's excellence in capturing the relationships between variables and accurately predicting outcomes. In the context of web development, the Decision Tree model was successfully utilised to fulfil the requirements of the project.

3.6 Random Forest Regression

Ensemble learning, specifically the creation of numerous decision trees, plays a pivotal role in both classification and regression tasks. Decision trees are fundamental components in various machine learning problems, offering versatile methodologies for tackling diverse challenges. The inherent scalability and robustness of tree-based learning make it an indispensable tool for data mining, capable of withstanding scaling and other transformations. To capture intricate patterns effectively, decision trees are often grown to a considerable depth. Random forest, a popular ensemble learning technique, leverages the aggregation of multiple deep decision trees trained on different subsets of the same training set. While this approach introduces a minor increase in bias and some trade-offs in interpretability, it enables improved performance through the ensemble's collective predictive capabilities.

3.6 Support Vector Regression

Supervised learning encompasses a family of learning algorithms that systematically analyse data to facilitate classification and regression analysis. These algorithms leverage labelled training data to establish patterns and relationships, enabling accurate predictions and insights.

4. Project

In this research paper, we present the development of a house price estimation model specifically designed for properties in Pune. The model is constructed using machine learning techniques, allowing for accurate predictions of house prices based on various relevant features. Furthermore, we deploy the model on AWS's Elastic Beanstalk platform, ensuring its availability and Accessibility to users. The aim of this research is to provide an effective tool for estimating house prices in Pune, leveraging the power of machine learning and cloud-based deployment for real-time predictions.

4.2 Data

The significance of data in a machine learning assignment cannot be overstated, as it forms the foundation upon which accurate and reliable findings are built. The origin, presentation, consistency, presence of outliers and other

characteristics of the data heavily influence the outcomes of the analysis. Therefore, careful attention must be given to addressing numerous questions to ensure the effectiveness and correctness of the learning algorithm. This involves undertaking several sub-steps, such as data acquisition, cleaning, and conversion. In this research paper, we delve into these crucial steps to comprehensively understand how they have been implemented in our project and the inherent value they bring to the machine learning component.

4.3 Data Set

<https://www.kaggle.com/datasets/levanshbhan/pune-house-data>

4.4 Model

https://github.com/LevanshBhan/Pune_House_Price_Prediction/blob/main/Pune_Housing_Price_Prediction.ipynb

4.5 Project Architecture

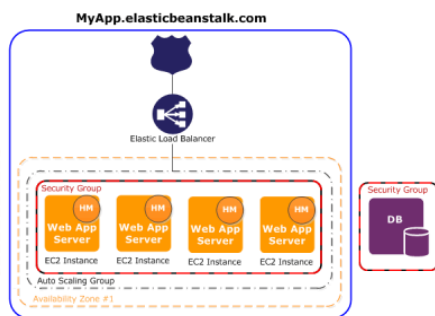


Figure 1: Architecture of the Application

4.6 Data Science

The first stage involves standard data science work, where we acquire the dataset named 'Pune House pricing data' from a reliable source such as Kaggle. To ensure reliable predictions throughout the prediction phase, we perform extensive data cleaning on the dataset. Our data science tasks are conducted within the Jupyter notebook titled 'Pune-HousePrice-Prediction-Model.ipynb'. While the notebook itself provides self-explanatory insights, we briefly touch upon the implemented principles. Data cleansing requires significant effort, with approximately 70% of the notebook dedicated to this process. It involves removing empty rows and eliminating irrelevant columns that do not contribute to the prediction task.

The next stage focuses on extracting valuable and meaningful information from the dataset, as it plays a crucial role in achieving accurate predictions. This stage involves identifying the most relevant features and optimizing the dataset for prediction.

In the final stage, we address the challenge of outliers, which can have a significant impact on data integrity and prediction accuracy. Understanding the dataset and detecting outliers allows us to effectively manage their influence.

By completing these stages, we transform the original dataset, which initially consists of over 13,000 rows and 9 columns, into a refined version with approximately 7,000 rows and 5 columns. This reduction ensures a streamlined and focused dataset, facilitating further analysis and prediction tasks.

4.7 Machine Learning

The preprocessed data is then utilised to train a machine-learning model. In order to determine the optimal procedure and parameters for the model, we primarily employ techniques such as K-fold Cross-Validation and the GridSearchCV approach. Through these methods, we aim to enhance the performance and generalisability of the model.

Upon experimentation, it is revealed that the linear regression model yields the most favourable outcomes for our dataset, achieving a score of over 80%, which is reasonably promising.

To ensure accessibility and ease of deployment, we export the trained model as a pickle file named 'Pune_House_Data.pickle'. This file format allows Python objects to be transformed into a serialized character stream. Additionally, to facilitate interaction with the model's features (columns) from the front end, we export them into a JSON file named 'columns.json'. This format enables seamless integration and utilization of the model in practical applications.

4.8 Frontend

The front end of the system is developed using simple HTML, providing a user-friendly interface. To obtain an estimated house price, users can conveniently input the relevant details such as the number of square feet, BHK (bedrooms, hall, kitchen), bathrooms, and location into a form. By clicking the 'ESTIMATE PRICE' button, the form data is submitted.

To handle the backend functionality, we utilize Flask Server, which is configured in Python. The server receives the form data entered by the user and executes the corresponding function. This function utilizes the trained prediction model to calculate the projected house price in lakhs of rupees. It efficiently processes the input data and provides an accurate estimation based on the model's predictions. It is important to note that in this context, 1 lakh is equivalent to 100,000 rupees.

5. EXPERIMENTAL SETUP

5.1 Steps to Create Model

1. Library Importation
2. Dataset Loading
3. Analysis of Data Exploratory
4. Data Cleansing
5. Feature Engineering
6. Dimensionality Reduction
7. Outlier Elimination based on Business Logic
8. Outlier Elimination using Standard Deviation and Mean
9. Data Visualisation
10. Model Construction
11. Model Testing on Selected Properties
12. Export the Tested Model to a Pickle File

5.2 Steps to Deploy Model on AWS Using Amazon's Elastic Beanstalk

1. Access your AWS console and locate Elastic Beanstalk.
2. Click on the "Create application" option and proceed to create it.
3. Click the "Create environment" button.
4. Create the environment.
5. In the Environment tier, verify the Web server information.
6. In the Environment information section, modify the Domain as necessary.
7. Choose Python as the platform and maintain the default settings for the Platform branch and version.
8. Specify the appropriate options in the Application code section, ensuring to provide a unique name for the Version label.
9. Utilise the "Choose file" option to upload the zipped file.
10. Select the "Single instance (free tier)" option for the Configuration presets.
11. Proceed to the next step.

12. Leave the defaults in the Configure service access section.
13. Move to the next step.
14. Confirm that "Activated" is selected in the Instance settings, then proceed to the next step.
15. If applicable, complete any optional sections by selecting the "Next" option.
16. Click on the "Submit" button to finalise the process.

5.3 Tools Used

1. Anaconda
2. Jupyter Notebook
3. Google Colaboratory 4
4. Flask
5. Amazon Web Services
6. EC2
7. Elastic Beanstalk

5.4 Technologies Used

1. Python
2. HTML
3. CSS
4. Bootstrap

6. Results




Figure 2: User Interface

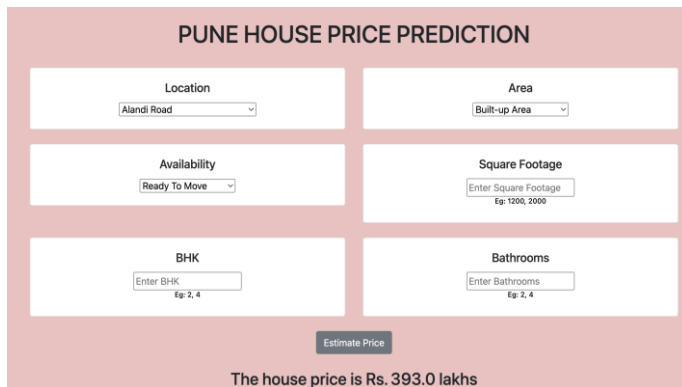


Figure 3: Predicting Price

AWS Elastic Beanstalk Web Application

6.1 Conclusion

In conclusion, this research project successfully navigated the intricate intersection of machine learning, data science, and web development to present a robust model for predicting house prices in Pune, India. Leveraging techniques such as linear regression and deploying the model on AWS Elastic Beanstalk, we achieved promising accuracy, surpassing 80%. The integration of user-friendly interfaces and efficient deployment mechanisms not only enhances accessibility but also underscores the practicality of our approach.

This endeavour not only fortified our skills in machine learning but also contributes a valuable tool to the real estate domain, fostering advancements in predictive analytics and offering a tangible solution to the challenges of price fluctuations in the housing market.

REFERENCES

1. Model "PUNE HOUSE PRICE PREDICTION MODEL"
2. Amazon's Elastic Beanstalk Documentation
3. Pickle Documentation
4. Repository
5. Grinberg, Miguel. Flask web development: developing web applications with Python. " O'Reilly Media, Inc.", 2018.
6. Aggarwal, Shalabh. Flask framework cookbook. Packt Publishing Ltd, 2014.
7. Musciano, Chuck, and Bill Kennedy. HTML & XHTML: The Definitive Guide: The Definitive Guide. " O'Reilly Media, Inc.", 2002.

8. A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1936-1939, doi: 10.1109/ICICCT.2018.8473231.
9. Real Estate Price Prediction with Regression and Classification. CS 229 autumn 2016 Project Final Report. Hijua Yu, Jiufa Wu.
10. House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. Visit Limsombunchai Commerce Division, Lincoln University. P. Acharjya: A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools.