

HOUSE PRICE PREDICTION BASED ON MACHINE LEARNING

Aditya wankhade¹, Umesh shinde², Hrushikesh Wayal³

Department of Computer Engineering, University of Pune, Pune, MH, India

Abstract — We propose to implement a house price prediction model of Bangalore, India. It's a Machine Learning model which integrates Data Science and Web Development. We have deployed the app on the Heroku Cloud Application Platform. Housing prices fluctuate on a daily basis and are sometimes exaggerated rather than based on worth. The major focus of this project is on predicting home prices using genuine factors. Here, we intend to base an evaluation on every basic criterion that is taken into account when establishing the pricing. The goal of this project is to learn Python and get experience in Data Analytics, Machine Learning, and AI.

Keywords — Bangalore House Price Prediction (BHPP), Pandas, NumPy, Matplotlib, Exploratory Data Analysis (EDA), Data Cleaning, Feature Engineering, Dimensionality Reduction, Data Visualization, Heroku.

1. INTRODUCTION

1.1 Context

This project was made because we were intrigued and we wanted to gain hands-on experience with the Machine Learning Project.

1.2 Motivation

We are highly interested in anything related to Machine Learning, the independent project provided us with the opportunity to study and reaffirm our passion for this subject. The capacity to generate guesses, forecasts, and offer machines the ability to learn on their own is both powerful and infinite in terms of application possibilities. Machine Learning may be applied in finance, medicine, and virtually any other field. That is why we opted to base our idea on Machine Learning.

1.3 Objective

As a first project, we intended to make it as instructional as possible by tackling each stage of the machine learning process and attempting to comprehend it well. We have picked Bangalore Real Estate Prediction as a method, which is known as a "toy issue," identifying problems that are not of immediate scientific relevance but are helpful to demonstrate and practice. The objective was to forecast the price of a specific apartment based on market pricing while accounting for various "features" that would be established in the following sections.

2. LITERATURE SURVEY

Real Estate Property is not only a person's primary desire, but it also reflects a person's wealth and prestige in today's society. Real estate investment typically appears to be lucrative since property values do not drop in a choppy fashion. Changes in the value of the real estate will have an impact on many home investors, bankers, policymakers, and others. Real estate investing appears to be a tempting option for investors. As a result, anticipating the important estate price is an essential economic indicator. According to the 2011 census, the Asian country ranks second in the world in terms of the number of households, with a total of 24.67 crores. However, previous recessions have demonstrated that real estate costs cannot be seen. The expenses of significant estate property are linked to the state's economic situation. Regardless, we don't have accurate standardized approaches to live the significant estate property values.

First, we looked at different articles and discussions about machine learning for housing price prediction. The title of the article is house price prediction, and it is based on machine learning and neural networks. The publication's description is minimal error and the highest accuracy. The aforementioned title of the paper is Hedonic models based on price data from Belfast infer that submarkets and residential valuation this model is used to identify over a larger spatial scale and implications for the evaluation process related to the selection of comparable evidence and the quality of variables that the values may require. Understanding current developments in house prices and homeownership are the subject of the study. In this article, they utilized a feedback mechanism or social pandemic that fosters a perception of property as an essential market investment.

3. METHODOLOGY

3.1 Data Collection

The statistics were gathered from Bangalore home prices. The information includes many variables such as area type, availability, location, BHK, society, total square feet, bathrooms, and balconies.

3.2 Linear Regression

Linear regression is a supervised learning technique. It is responsible for predicting the value of a dependent variable (Y) based on a given independent variable (X). It

is the connection between the input (X) and the output (Y). It is one of the most well-known and well-understood machine learning algorithms. Simple linear regression, ordinary least squares, Gradient Descent, and Regularization are the linear regression models.

3.3 Decision Tree Regression

It is an object that trains a tree-structured model to predict data in the future in order to provide meaningful continuous output. The core principles of decision trees, Maximizing Information Gain, Classification trees, and Regression trees are the processes involved in decision tree regression. The essential notion of decision trees is that they are built via recursive partitioning. Each node can be divided into child nodes, beginning with the root node, which is known as the parent node. These nodes have the potential to become the parent nodes of their resulting offspring nodes. The nodes at the informative features are specified as the maximizing information gain, to establish an objective function that is to optimize the tree learning method.

3.4 Classification Trees

Classification trees are used to forecast the object into classes of a categorical dependent variable based on one or more predictor variables.

3.5 Regression Trees

It supports both continuous and categorical input variables. Regression trees are regarded as research with various machine algorithms for the regression issue, with the Decision Tree approach providing the lowest loss. The R-Squared value for the Decision Tree is 0.998, indicating that it is an excellent model. The Decision Tree was used to complete the web development.

3.6 Support Vector Regression

Supervised learning is linked with learning algorithms that examine data for classification and regression analysis.

3.7 Random Forest Regression

It is an essential learning approach for classification and regression to create a large number of decision trees. Preliminaries of decision trees are common approaches for a variety of machine learning problems. Tree learning is required for serving n off the self-produce for data mining since it is invariant despite scaling and several other changes. The trees are grown very deep in order to learn a high regular pattern. Random forest is a method of averaging several deep decision trees trained on various portions of the same training set. This comes at the price of a slight increase in bias and some interoperability.

4. PROJECT

4.1 Problem Statement

Create a model to estimate the price of houses in Bengaluru and host it on Heroku.

4.2 Data

The data is the most important aspect of a machine learning assignment, to which special attention should be paid. Indeed, the data will heavily affect the findings depending on where we found them, how they are presented, if they are consistent, if there is an outlier, and so on. Many questions must be addressed at this stage to ensure that the learning algorithm is efficient and correct.

To obtain, clean, and convert the data, many sub steps are required. We will go through these steps to understand how they've been used in my project and why they're helpful for the machine learning section.

4.3 Dataset

Dataset:

<https://www.kaggle.com/ameythakur20/bangalore-house-prices>

4.4 Model

Model:

<https://www.kaggle.com/ameythakur20/bangalore-house-price-prediction-model>

4.5 Data Science

The first stage is standard data science work, in which we take a data set named 'Bengaluru House pricing data' from Kaggle. We will do significant data cleaning on it to guarantee that it provides reliable predictions throughout prediction. This Jupyter notebook, 'Bangalore-House-Price-Prediction-Model.ipynb,' is where we do all of our data science work. Because the Jupyter notebook is self-explanatory, we will only touch on the principles that we have implemented briefly. In terms of data cleansing, our dataset needs a significant amount of effort. In fact, 70% of the notebook is dedicated to data cleaning, in which we eliminate empty rows and remove superfluous columns that will not aid in prediction.

The process of obtaining valuable and significant information from a dataset that will contribute the most to a successful prediction is the next stage.

The final stage is to deal with outliers. Outliers are abnormalities that do massive damage to data and prediction. There is a lot to comprehend conceptually from the dataset in order to discover and eliminate these outliers.

Finally, the original dataset of over 13000 rows and 9 columns is reduced to about 7000 rows and 5 columns.

4.6 Machine Learning

The resulting data is fed into a machine learning model. To find the optimal procedure and parameters for the model, we will mostly employ K-fold Cross-Validation and the GridSearchCV approach.

It turns out that the linear regression model produces the best results for our data, with a score of more than 80%, which is not terrible.

Now, we need to export our model as a pickle file (Bengaluru_House_Data.pickle), which transforms Python objects into a character stream. Also, in order to interact with the locations(columns) from the frontend, we must export them into a JSON (columns.json) file.

4.7 Frontend

The front end is built up of straightforward HTML. To receive an estimated pricing, the user may fill-up the form with the number of square feet, BHK, bathrooms, and location and click the 'ESTIMATE PRICE' button. We have used Flask Server and configured it in python. It takes the

form data entered by the user and executes the function, which employs the prediction model to calculate the projected price in lakhs of rupees (1 lakh = 100000).

5. EXPERIMENTAL SETUP

5.1 Steps to Create Model

1. Import Libraries
2. Load Dataset
3. Exploratory Data Analysis
4. Data Cleaning
5. Feature Engineering
6. Dimensionality Reductions
7. Outlier Removal using Business Logic
8. Outlier Removal using Standard Deviation & Mean
9. Data Visualization
10. Building a Model
11. Test the Model for few properties
12. Export the tested model to a pickle file

5.2 Tools used

1. Anaconda
2. Jupyter Notebook

5.3 Technologies used

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

6. RESULTS

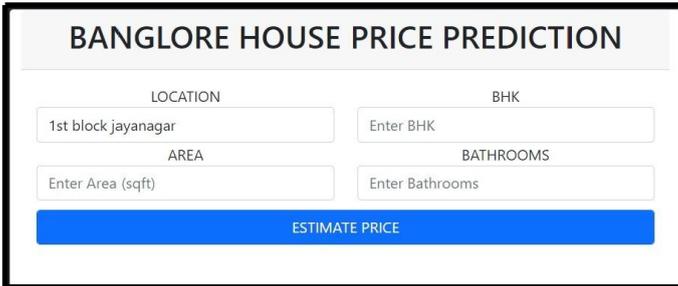


Figure 2: User Interface

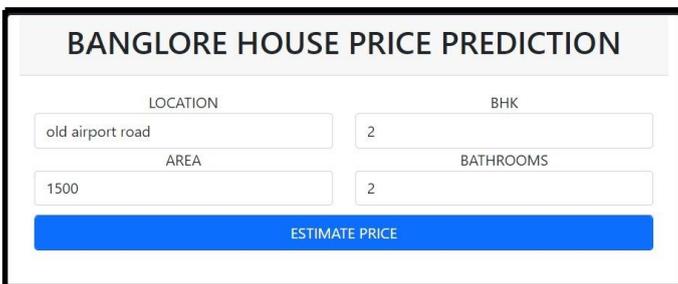


Figure 3: Estimate Price



Figure 4: Predict

CONCLUSIONS

With several characteristics, the suggested method predicts the property price in Bangalore. We experimented with different Machine Learning algorithms to get the best model. When compared to all other algorithms, the Decision Tree Algorithm achieved the lowest loss and the greatest R-squared. Flask was used to create the website.

Let's see how our project pans out. Open the HTML web page we generated and run the app.py file in the backend. Input the property's square footage, the number of bedrooms, the number of bathrooms, and the location, then click 'ESTIMATE PRICE.' We forecasted the cost of what may be someone's ideal home.

REFERENCES

- [1] Model "BANGALORE HOUSE PRICE PREDICTION MODEL"
- [2] Heroku "Documentation"
- [3] Repository: "Web Application" <https://github.com/msatmod/Bangalore-House-Price-Prediction>
- [4] Repository: "Web Application" <https://github.com/Amey-Thakur/BANGALORE-HOUSE-PRICE-PREDICTION>
- [5] Pickle "Documentation"
- [6] 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.
- [7] Furia, Palak, and Anand Khandare. "Real Estate Price Prediction Using Machine Learning Algorithm." e-Conference on Data Science and Intelligent Computing. 2020.
- [8] Musciano, Chuck, and Bill Kennedy. HTML & XHTML: The Definitive Guide: The Definitive Guide. " O'Reilly Media, Inc.", 2002.
- [9] Aggarwal, Shalabh. Flask framework cookbook. Packt Publishing Ltd, 2014.
- [10] Grinberg, Miguel. Flask web development: developing web applications with python. " O'Reilly Media, Inc.", 2018.
- [11] Middleton, Neil, and Richard Schneeman. Heroku: up and running: effortless application deployment and scaling. " O'Reilly Media, Inc.", 2013.