

HOUSE PRICE ESTIMATION USING DATA SCIENCE AND ML

Sai Bhavan Gubbala¹, Dhruv Naidu Alti², Leela Dhanush Naidu Kalaparathi³, Adharsh Boora⁴

¹² Student, Dept. of CSE, Sathyabama Institute of Science and Technology, Chennai India.

³ Student, Dept. of Mechanical, Sathyabama Institute of Science and Technology, Chennai India.

⁴ Student, Dept. of ECE, Sathyabama Institute of Science and Technology, Chennai India.

Abstract - We propose implementing a property price forecast model for Bangalore, India. It is a Machine Learning paradigm that combines Data Science with Web Development. Because housing prices are significantly connected with other characteristics such as location, region, and population, predicting individual house prices requires information other than the HPI. Many articles employ typical machine learning algorithms to estimate house prices reliably. However, they seldom concern themselves with the performance of individual models and ignore the less popular yet complicated models. As a result, housing prices change daily and are frequently inflated rather than based on value. The preeminent goal of this research is to forecast property values using real-world criteria. Here, we evaluate each primary factor considered when determining price. In addition, this project aims to study Python and gain expertise in Big Data, Machine Learning, and AI.

Key Words: Machine Learning, HPI, Big Data, Artificial Intelligence, Data Science, Prediction.

1. INTRODUCTION

Both human living standards and global economic growth are correlated with the level of house prices. Over the last Six and a half years, Bangalore's residential real estate market has witnessed a phenomenal exchange in the housing stock. However, the market, formerly thought to be one of the most substantial private marketplaces in the country, is currently reeling from fear and struggling to stay afloat during challenging circumstances.

1.1 MICROSOFT POWER BI

Despite the company's size, Power BI is a cloud-based product requiring no infrastructure maintenance or financial investment. The tool's current iteration is unrestricted by old software, and its users do not require special training to provide business intelligence insights. In addition, the installation of Power BI embedded is simple and quick, as with other Microsoft cloud services.

1.2 PYTHON

It is an advanced programming language intended to be simple to understand and implement. It is open-source software, which implies that it may be used for free, even in commercial applications. Python is available for Mac,

Windows, and Unix computers, and it has also been adapted to Java and .NET virtual machines. It is also supported by various 2D and 3D imaging tools, allowing users to utilize Python to construct custom plug-ins and extensions.

2. LITERATURE SURVEY

The associated work survey aims to focus on and discover the best functioning methodology and strategies for determining the most desirable, reasonable, and cheap housing price forecasting study. Property not only constitutes a person's primary goal but also indicates a person's status and wealth in today's society. Real estate investing looks profitable since house prices do not collapse jaggedly. Real estate value fluctuations will affect many home stakeholders, bankers, legislators, and others. Real estate is an appealing alternative for investors. As a result, forecasting the significant estate price is a critical economic indicator. According to the estimated population, the Asian country ranks top 10 in terms of the household population, with a total of 27.42 crores. However, prior recessions have shown that real estate expenses are not visible. Significant estate property costs are tied to the state's economic status. Regardless, we lack precise standardized methods for living the significant estate property values.

According to the data, the Random Forest Regressor gave the highest accuracy, trailed by the Decision Tree Regression model. Random Forest produces comparable results, with a minimal reduction in Lasso. There is no dramatic difference between all feature selection groups, independent of positive or negative groups. It is a positive indication that the purchase prices may be used only to forecast the selling prices without considering additional factors to reduce model over-fitting. Furthermore, there is a decrease in accuracy in the fragile features group. The Root Square Mean error shows the same pattern of results for all feature choices.

3. DATA SET INFORMATION

The data is the most crucial part of a machine-learning project and should be carefully considered. Indeed, the data will significantly impact the conclusions depending on where we obtained them, how they are presented, if they are consistent, whether there is an outlier, and so on. At this step, several questions must be answered to guarantee that the learning algorithm is effective and correct. The collection includes data on 13320 instances and nine characteristics. In

addition, the following characteristics are presented: area type, availability, location, size, society, total sqft, bath, balcony, and price after removing all null values from the dataset. The dataset has been completed and is available for manipulation.

4.FLOWCHART OF DIAGRAM

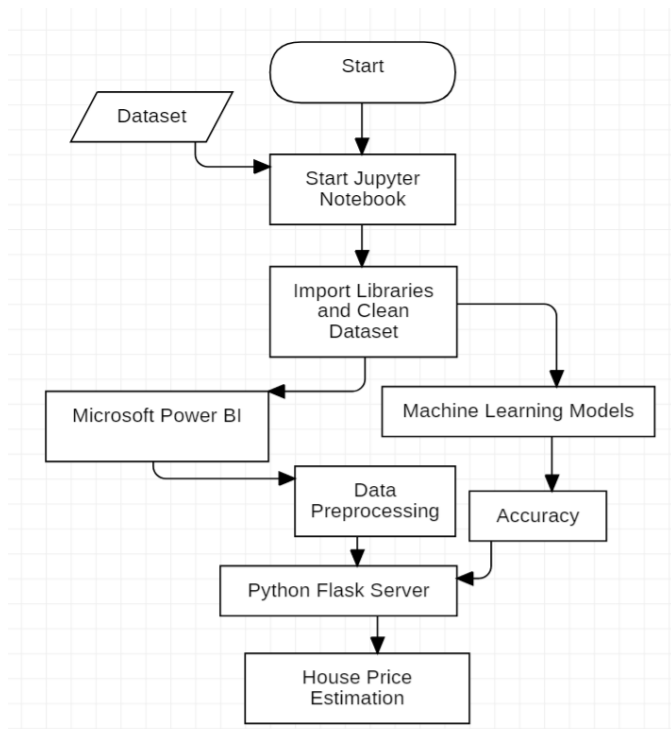


Fig -1: System Architecture

4.1 DATASET PREPROCESSING

The figures below show that the data contains some null values that Null values will randomly adjust, and the dataset will thoroughly clean the information.

```

df2.isnull().sum()
location    1
size       16
total_sqft  0
bath       73
price      0
dtype: int64

df3 = df2.dropna()
df3.isnull().sum()
location    0
size       0
total_sqft  0
bath       0
price      0
dtype: int64
  
```

Fig -2: Cleaning Dataset



Fig -3: Attributes Comparison with Price

4.2 Machine Learning Models

```

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)
  
```

Fig -4: Model Training

5.METHODOLOGY

The processed data with the highest accuracy will be picked for house price estimation, and the forecast process on the web application will be monitored. Python will execute the flask server in Microsoft Visual Studio Code, and a local server will be enabled to evaluate home prices in a particular location.

5.1 LINEAR REGRESSION

Linear regression aims to predict the relationship between two variables by fitting a equation to observable data. One variable is considered an explanatory variable, while the other is considered a dependent variable. A modeler, for example, could wish to apply a linear regression model to match people's pounds to their measurements.

5.2 LASSO

The LASSO approach regularizes model parameters by decreasing part of the regression coefficients to zero. Following the shrinkage, the feature selection phase follows, in which every non-zero value is chosen to be incorporated into the model. This strategy helps reduce prediction mistakes that are typical in statistical models.

5.3 DECISION TREE CLASSIFIER

Decision tree learning applies a divide-and-conquer technique by undertaking a greedy search to determine the best split points inside a tree. This dividing procedure is then continued in a top-down, recursive way until all or the majority of records have been categorized under particular class labels. The decision tree's complexity determines whether or not all data points are classified as homogeneous sets.

5.4 OUTLIERS

The minimum price per square foot is 263 Rs/sqft, and the maximum is 13 million, indicating a significant range in property values. Therefore, there is a need to use the mean and one standard deviation to eliminate outliers from each location. Some of the outliers of Hebbal and Rajaji street can be see below. These outliers are removed after cleaning the dataset. Based on the charts below, we can see that the data points indicated in red are outliers and are being deleted by the remove BHK outliers function.

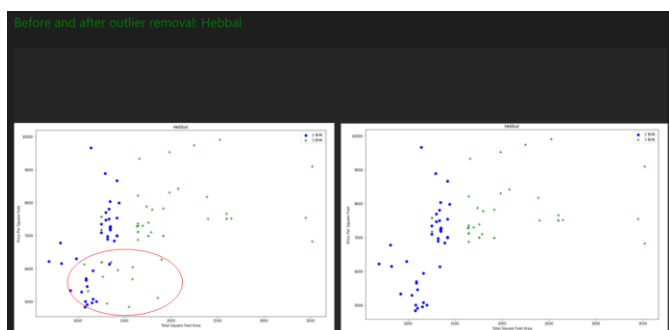


Fig -5: Hebbal Outliers

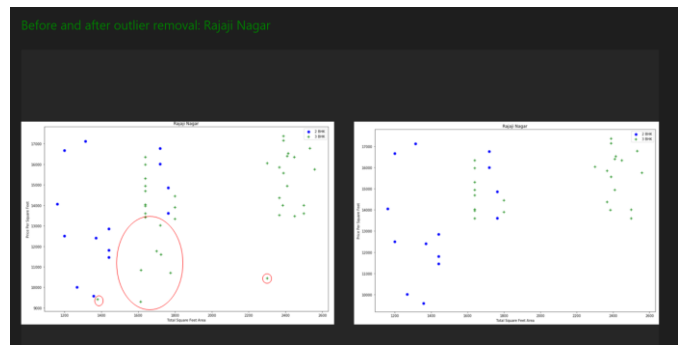


Fig -6: Rajaji Nagar Outliers

5.5 IMPORT PICKLE AND LOCATION INFORMATION

Pickle is used in Python to serialize and deserialize a Python object structure. In other words, it transforms a Python object into a byte stream to save it to a file/database, retain a program state between sessions, or transport data over a network. The original object hierarchy may be recreated by unpickling the pickled byte stream. This entire procedure is comparable to object serialization in Java or .NET.

In the Python script or module, a JSON package must be imported if the data needs to be serialized or deserialized. JSON employs comma-separated key-value pairs enclosed in double quotation marks and separated by colons. In addition, curly braces {} or square brackets [] (also called "brackets" in certain countries) can be used to delimit the body of a JSON file. The JSON format looks to be similar to the Python dictionary, but the intricacies of the JSON format differ significantly; therefore, take caution while working with both forms.

```
import pickle
with open('bangalore_bhavan_home_prices_model.pickle','wb') as f:
    pickle.dump(lr_clf,f)
```

Fig -7: Importing Pickle

```
import json
columns = {
    'data_columns': [col.lower() for col in X.columns]
}
with open("columns.json", "w") as f:
    f.write(json.dumps(columns))
```

Fig -8: Importing JSON

6.RESULTS AND CONCLUSION

According to the results, the Linear Regression Model obtained the most remarkable accuracy, whereas the other two algorithms produced comparable accuracy lower than

the higher-attaining accuracy. Finally, a linear regression model will be implemented in the web application to supervise the estimating process and calculate the price of the building in that specific location. Linear Regression has achieved an accuracy of 84.77.

	model	best_score	best_params
0	linear_regression	0.847796	{'normalize': False}
1	lasso	0.726748	{'alpha': 2, 'selection': 'random'}
2	decision_tree	0.713610	{'criterion': 'friedman_mse', 'splitter': 'best'}

Fig -9: Accuracy of the models

After exporting the necessary files, we created a web application in Flask that allows users to enter attributes and obtain an estimated price for a home or flat in Bangalore. The graphics below depict the results of our investigation.

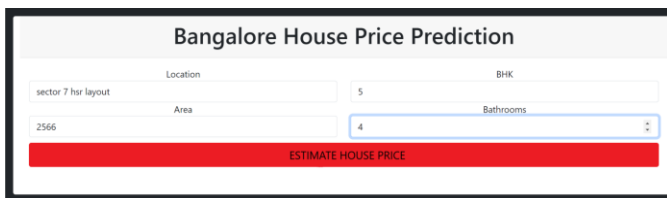


Fig -10: Web Application Interface

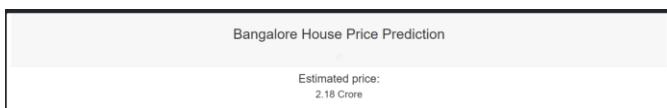


Fig -11: Price Estimation

7.Future Work

People will be able to utilize this program in the future to acquire the most accurate pricing of a home. This application may be converted into a Flutter application to get support for Android and iOS devices, allowing it to be used everywhere. It can also be used as an external or internal service for apps that display property for rent. Users may apply this methodology to various fields, such as tuition costs in a specific location, swimming pool rates, and data science-type models.

REFERENCES

- [1] <https://www.kaggle.com/datasets/amitabhajoy/bengaluru-house-price-data>
- [2] Thakur, Amey & Satish, Mega. (2021). Bangalore House Price Prediction. 8. 193-196.
- [3] Manasa, J & Gupta, Radha & Nuggenahalli, Narahari. (2020). Machine Learning based Predicting House Prices using Regression Techniques. 624-630. 10.1109/ICIMIA48430.2020.9074952.
- [4] Sinha, Anurag & Ramish, Md. (2021). HOUSE COST ESTIMATION OF BANGALORE REGION USING FEATURE SELECTION ALGORITHM OF MACHINE LEARNING.
- [5] Truong, Quang & Nguyen, Minh & Dang, Hy & Mei, Bo. (2020). Housing Price Prediction via Improved Machine Learning Techniques. Procedia Computer Science. 174. 433-442. 10.1016/j.procs.2020.06.111.
- [6] Zulkifley, Nor & Rahman, Shuzlina & Nor Hasbiah, Ubaidullah & Ibrahim, Ismail. (2020). House Price Prediction using a Machine Learning Model: A Survey of Literature. International Journal of Modern Education and Computer Science. 12. 46-54. 10.5815/ijmecs.2020.06.04.
- [7] Deo, Udit. (2021). House Price Prediction. 10.13140/RG.2.2.27657.98408.
- [8] Kang, Yuhao & Zhang, Fan & Peng, Wenzhe & Gao, Song & Rao, Jinmeng & Duarte, Fábio & Ratti, Carlo. (2020). Understanding house price appreciation using multi-source big geo-data and machine learning. Land Use Policy. 111. 10.1016/j.landusepol.2020.104919.
- [9] Özdemir, Ozancan. (2022). House Price Prediction Using Machine Learning: A Case in Iowa. 10.13140/RG.2.2.19846.86086.
- [10] Hannonen, Marko. (2020). A New Methodology for House Price Analysis.
- [11] Browning, Martin & Gortz, Mette & Leth-Petersen, Søren. (2011). House Prices and Consumption: A Micro Study.
- [12] Fernandez-Duran, Laura & Llorca, Alicia & Ruiz, N & Valero, S. & Botti, V.. (2011). The impact of location on housing prices: applying the Artificial Neural Network Model as an analytical tool. ERS conference papers.
- [13] Lee, Min-feng & Chen, Guey-shya & Lin, Shao-pin & Wang, Wei-jie. (2022). A Data Mining Study on House Price in Central Regions of Taiwan Using Education Categorical Data, Environmental Indicators, and House Features Data. Sustainability. 14. 6433. 10.3390/su14116433.
- [14] Kholodilin, Konstantin & Siliverstovs, Boriss & Menz, Jan-Oliver. (2007). What Drives Housing Prices Down? Evidence from an International Panel. Journal of Economics and Statistics (Jahrbuecher fuer Nationaloekonomie und Statistik). 230. 59-76. 10.1515/jbnst-2010-0105.
- [15] Zietz, Joachim & Zietz, Emily & Sirmans, G.. (2008). Determinants of House Prices: A Quantile Regression Approach. The Journal of Real Estate Finance and Economics. 37. 317-333. 10.1007/s11146-007-9053-7.

- [16] Wu, Lynn & Brynjolfsson, Erik. (2013). The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. SSRN Electronic Journal. 10.2139/ssrn.2022293.