HOUSING PRICE PREDICTION

Sricharan Gudi
Department of CSE,
SRM University-AP, India,
sricharan_gudi@srmap.edu.in

Mounish Sai Mididodla
Department of CSE,
SRM University-AP, India,
mounishsai_m@srmap.edu.in

SaiPrakash Paladugu
Department of CSE,
SRM University-AP, India,
saiprakash_p@srmap.edu.in

Gopal Lanka
Department of CSE,
SRM University-AP, India,
gopal_lanka@srmap.edu.in

Sumalatha Saleti
Assistant Professor,
Department of CSE,
SRM University-AP, India,
sumalatha.s@srmap.edu.in

Abstract

In most cases, the House price index indicates the aggregate amount of residential living price fluctuations. To make it easier for a family to choose a residence, we have tailored it more specifically via inquiries for the needed square feet, number of bedrooms, and restrooms. This research investigates a realistic data pre-processing, creative feature engineering strategy using preloaded datasets and data features. In addition, the research provides a regression approach in machine learning to estimate property prices.

I. Introduction

Data has become the heart of technological advancements, and forecasting algorithms may now achieve any result. This method heavily relies on machine learning. Machine learning is about providing a reliable dataset and then making projections based on it. The machine understands how important a specific event is to the overall system based on its pre-loaded data and predicts the outcome appropriately [1].

Machine learning has evolved into an essential prediction technique in recent times, owing to the increasing shift towards Big Data, because it can estimate property values more correctly based on their qualities, regardless of previous year's data. Several research investigated this issue and shown the potential of the machine learning technique. Instead of considering into factor the combination of many machine learning models, the majority of them just assessed the models' respective performances [2].

Housing prices are a key economic indicator, while price ranges are significant to consumers as well as sellers. House prices will be forecasted in this study using explanatory factors that encompass several elements of residential properties. On constant home prices, they will be forecasted using linear regression techniques such as K-fold. Our primary objective is to forecast a property price based on their requirements and goals. Future prices will be projected by analysing recent market patterns and price ranges, as well as prospective changes [3].

II. Literature Review

In this article [4], we'll go over our answer for the "House Prices: Advanced Regression Techniques" machine learning rivalry, which was hosted on the Kaggle platform. The purpose is to forecast house selling prices based on factors such as house size, year of construction, and so on. We employ standard machine learning techniques as well as our own methods, which are discussed here. We finished 18th out of 2124 competitors from all around the world in the highest level of the tournament.

House price forecasting is an important aspect of real estate. The literature seeks to glean relevant information from historical property market data. Strategies based on machine learning are used to analyse previous real estate transactions in Australia in order to create helpful models for home buyers and sellers. The wide disparity in housing prices between Melbourne's most expensive and least expensive areas has been revealed. Furthermore, investigations show that a combination of Stepwise and Support Vector Machine with mean squared error assessment may be a competitive method.[5]

The housing sector is one of the most price-sensitive in the world, and it is always changing. It is one of the most important domains for applying machine learning concepts to improve and forecast selling prices with high precision. Physical conditions, ideas, and location are three variables that impact the price of a property. The existing framework involves calculating the value of properties without regard for market prices or price increases. The goal of this article is to forecast home prices for buyers based on their financial goals and desires. Future costs will be predicted by breaking down historical market trends and value ranges, as well as upcoming developments. This investigation aims to forecast housing values in Mumbai using Linear Regression. It will assist clients in putting funds in donations without using a broker. This study found that linear regression has the lowest prediction error, which is 0.3713.[6]

In this work, we use text mining and machine learning to forecast Dutch housing trends as a form of knowledge science approaches in finance. Using text information collected from Twitter, we want to forecast the short-term increasing or negative trend of the average property value in the Dutch market. Twitter is very frequently utilised and has been shown to be a valuable source of information. However, Instagram, text analysis (tokenization, collection of phrases, n-grams, weighted term frequencies), and machine learning (classification algorithms) have yet to be coupled in order to anticipate short-term housing market movements. In this investigation, tweets containing predetermined search terms are gathered using expertise in the field, and the related content is then categorised monthly as documents. Words and word sequences are then converted into numerical numbers. These numbers were used to forecast whether the home market will rise or fall. We addressed this as a type of binomial classification issue, linking monthly text data to (up or down) patterns for the next month. Our major findings show that there is a link between the average (weighted) frequency of words and short-term housing trends; in a nutshell, we were able to produce accurate short-term forecasts using a combination of machine learning and text mining approaches.[7]

Real estate is the industry with the least transparency in our ecosystem. Housing prices fluctuate on a daily basis and are frequently exaggerated rather than supported by appraisal. The major focus of our scientific study is on predicting house prices using real-world facts. Here, we intend to base our evaluations on every essential criteria that is taken into account when assessing the worth. Throughout this process, we employ a variety of regression approaches, and the outcomes are based on the weighted average of many methodologies to provide the most accurate results. The findings demonstrated that this strategy produces less error and more accuracy than separate algorithms. We also recommend integrating real-time neighbourhood information from Google Maps to encourage precise real-world values.[8]

III. Dataset Description

We employ the Bengaluru house price dataset, which is made up of enlightened data. Initially this dataset contains 9 features and 13320 sold property prices. Despite being small, the dataset covers 9 characteristics including area type, availability, location, price, size, total square feet, society, balcony, and number of bathrooms. Because there are so many characteristics, we can explore with different algorithms to estimate house values. This report outlines the development of a machine learning model for predicting housing prices in Bengaluru. The work includes data loading, exploration, cleaning, feature engineering, outlier removal, dimensionality reduction, and model building. The primary goal is to create an accurate model that can predict housing prices based on various property features.

Data Preprocessing:

It is the transformation of raw, complicated data into organized, intelligible information. It will search the dataset for missing and redundant data. As a result, the dataset becomes more consistent.

i. Data Loading and Exploration:

The project begins by loading the dataset into a pandas data frame. Initial exploration involves examining the dataset's shape, columns, and unique values in the 'area_type' column. This step provides an understanding of the data structure and the types of features available.

ii. Feature Selection:

Features that are deemed unnecessary for building the model, such as 'area_type', 'society', 'balcony', and 'availability', are dropped from the dataset. This simplifies the data and focuses on relevant information for predicting housing prices.

iii. Data Cleaning:

Missing values in the dataset are handled by dropping rows with NaN(not a number) values. This ensures that the model is trained on complete and reliable data, preventing potential issues during analysis.

iv. Feature Engineering:

New features are introduced to enhance the model's predictive capabilities. The 'bhk' feature is created by extracting the number of bedrooms from the 'size' column. The 'total_sqft' feature is processed to handle different formats, such as ranges, using a conversion function.

$$total\_sqft = \frac{x+y}{2}$$

Additionally, a 'price_per_sqft' feature is added to represent the cost per square foot of properties.

$$price - per - sqft = \frac{price*100000}{total\_sqft}$$

v. Dimensionality Reduction:

To simplify the dataset and reduce the number of location categories, locations with fewer than 10 data points are labelled as 'other'. One-hot encoding is then applied to the 'location' column to create dummy variables. This reduces the dimensionality of the dataset, making it more manageable for modelling.

vi. Outlier Removal:

Outliers are identified and removed based on business logic. For example, consider a sqft per bedroom is 300, Now if we take 2bhk the sqft would be 600 but in some cases, we have an sqft of 400 for a 2 bhk apartment, so this can be considered as an outlier. To remove this we have considered a threshold of 300 sqft_per_bhk, properties with less than 300 sqft per bedroom are considered outliers. Now we have considered the dataset by ignoring the outliers as discussed above.

Outliers in the 'price_per_sqft' column are also addressed using standard deviation and mean values. This step ensures that the model is not influenced by extreme data points. To achieve this we calculated the mean & standard deviation price_per_sqft column and modified the data frame where we included only the rows where price_per_sqft falls within the range of (mean - standard deviation) and (mean + standard deviation).

Now from the dataset we have considered the locations Rajaji Nagar and Hebbal to check how 2bhk & 3bhk property prices look like:

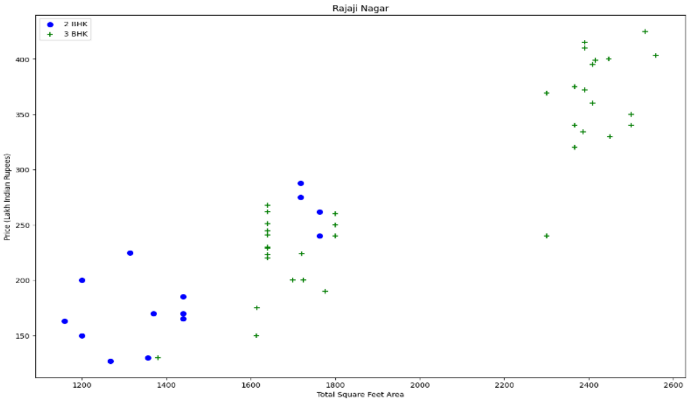

Fig 1: Observation of total_sqft vs price of Rajaji nagar before outlier removal.

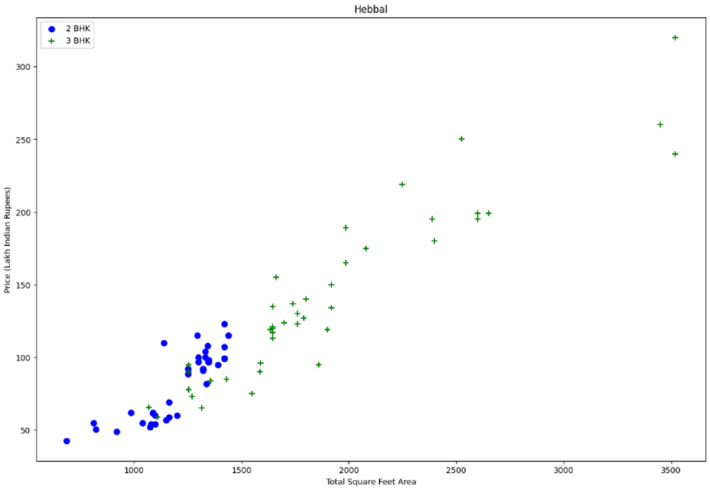

Fig 2: Observation of total_sqft vs price of Hebbal before outlier removal.

From fig1 & fig 2, we can observe for some of the 3bhk apartments the price is less than the 2bhk apartments. So, these are considered as outliers and removed those 2bhk apartments whose price_per_sqft is less than mean_price_per_sqft of 1bhk apartment rectified.

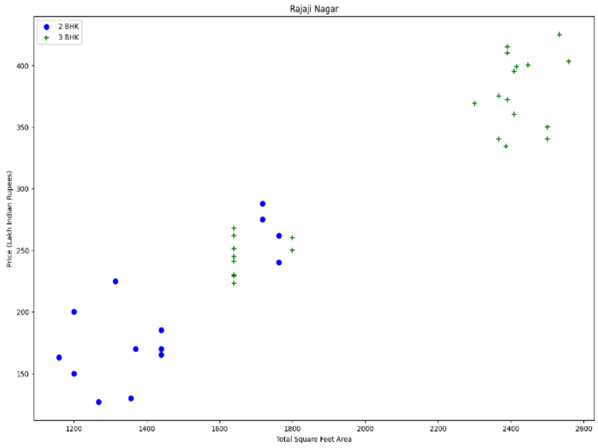

Fig 3: Rajaji Nagar after outlier removal.

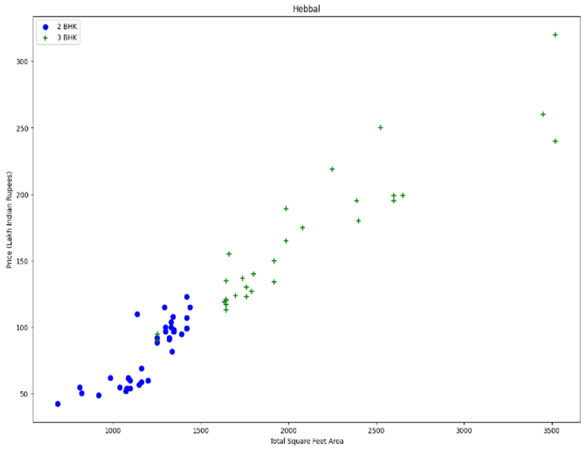

Fig 4: Hebbal after outlier removal.

Outlier removal using bathroom feature:

In some case we found that there are extra 2 bathrooms than no.of bedrooms, which we found as unusual. For example, from a businessman perspective if he have a n bedrooms there can be n+1 bathrooms as one bathroom is for guests. Anything greater than n+1 is considered as an outlier and rectified.

Model Building:

The dataset is split into features (X) and the target variable (y). A Linear Regression model is trained on the data. The model's performance is evaluated using the test set of 20% and training set of 80%, providing insights into its predictive accuracy.

Cross-Validation:

K-fold cross-validation is employed to assess the model's accuracy. The model consistently achieves a score above 80% in 5 iterations, indicating its robust performance. So the mean accuracy is 84%.

Hyperparameter Tuning:

GridSearchCV is utilized to find the best hyperparameters for different regression models, including Linear Regression, Lasso, and Decision Tree Regression. The model with the best cross-validated score is selected for further analysis.

Model Evaluation:

The chosen model is evaluated based on its performance in cross-validation. Linear Regression consistently yields the best results with an accuracy of greater than 80%, affirming its suitability for predicting housing prices in Bengaluru.

Testing the Model:

The model is tested using a custom function that takes input features (location, square footage, bathrooms, and bedrooms) and predicts the corresponding housing price. This demonstrates the practical application of the trained model.

IV. Implementation:

Code: https://github.com/SricharanGudi/Data_Mining_Project

V. Conclusion

The developed model successfully addresses data cleaning, feature engineering, outlier removal, and dimensionality reduction. The choice of Linear Regression as the predictive model is supported by cross-validation results. The model can serve as a valuable tool for predicting housing prices in Bengaluru, assisting stakeholders in making informed decisions in the real estate market. Continuous monitoring and updates to the model may be necessary to ensure its relevance and accuracy over time.

References

[1] *House Price Prediction using machine learning K Pavan,T Raghul.*

[2] *Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. Procedia Computer Science, 174, 433-442.*

[3] *Yu, H., & Wu, J. (2016). Real estate price prediction with regression and classification. CS229 (Machine Learning) Final Project Reports.*

[4] *Viktorovich, P. A., Aleksandrovich, P. V., Leopoldovich, K. I., & Vasilevna, P. I. (2018, August). Predicting sales prices of the houses using regression methods of machine learning. In 2018 3rd Russian-Pacific conference on computer technology and applications (RPC) (pp. 1-5). IEEE.*

[5] *Phan, T. D. (2018, December). Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. In 2018 International conference on machine learning and data engineering (iCMLDE) (pp. 35-42). IEEE.*

[6] Ghosalkar, N. N., & Dhage, S. N. (2018, August). Real estate value prediction using linear regression. In *2018 fourth international conference on computing communication control and automation (ICCUBEA)* (pp. 1-5). IEEE.

[7] *Velthorst, M., & Güven, Ç. (2019, June). Predicting Housing Market Trends Using Twitter Data. In 2019 6th Swiss Conference on Data Science (SDS) (pp. 113-118). IEEE.*

[8] House Price Prediction Using Machine Learning and Neural Networks, Ayush *Varma ; Abhijit Sarma ; Sagar Doshi ; Rohini Nair.*