House Market Prices Predicting With Machine Learning

# Introduction

The global housing market was one of the biggest financial markets in the world. It is the main criteria for any individual or business. Anyone wants to buy a home, and any companies need to have office space.

Individuals always required to buy a house for their needs. Everyone thinks to need to have every comfort like a place, construction, bed-rooms, approvals at the correct price.

### Problem

Rather than other markets, housing markets are highly volatile sometimes, and people will pay to overprice for a house that not needed. Even non-popular will be overpriced rather than the central cities. If buyers think to look at previous prices before what they about to buy, they astounded.

Chart, bar chart

Description automatically generated

The above statistics show that the rural area has the highest price per sqft than other cities(like Bangalore, Mumbai). What factors were derived? Realtors were made these prices too high to gain profits even not much worth it.

### Hypothesis

Upon looking at these factors, if previous data has analysed, can we overcome these dead traps? Any Machine Learning model will help to predict the price if a user has mentioned certain features.

The total aim of the project will be able to help in buying the right home at the right price. For this project to succeed, specific techniques Machine Learning or Neural Networks used.

### Mind Map

## Progress To Date

## About

### Analysis

**Which cities will spend a high average on housing**

To Identify the above question. Two main tasks needed are needed. Identifying main cities in India and average spends on the housing market.

Chart, bar chart

Description automatically generated

India is the seventh-largest country in the world, Rather than Identifying the most spending housing areas. Another way is identifying main cities average spending will reveal which city spending most on real estate.

Agra, Chennai and Secunderabad commonly had IT industry, connecting highways across the country, and the cost of living here is very low compared to other cities.

**Which city was constructing real estate more rapidly**

Several factors needed to develop urban areas. Real estate plays a significant role in that. People across different parts of India will come to cities, finding for work & live here.

Chart, bar chart

Description automatically generated

Mumbai, Kolkata, Chennai, & Jaipur was rapidly developing on real estate. Mumbai, Kolkata, Chennai are coastal areas and most developed crowdsourced areas well. So that factors have influencing real estate development. The Prime minister has announced a high-speed railway and Delhi to Mumbai express highway connecting through Jaipur. These factors have influenced real estate to grow.

**How many bedrooms most people prefer**

Everyone will think before buying a home. How spacious it is and how many bedrooms are there, and many features are cross-checked before buying.

Chart, pie chart

Description automatically generated

Applying analysis on BHK Distribution percentage. Three bedrooms have 44.9%, two bedrooms have 37.8%, Four bedrooms has 9.8%, and One bedroom 5.1%. Most Middle-income person can mostly afford 2 or 3 bedrooms. So most housing categories mainly focus on two or three bedrooms. Upper and upper-middle-income groups can afford four and above.

**Did Who posted will influence the price.**

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Considering the data, the Owner has posted more amount than the dealer or builder. On contradictory Owner has posted more SQFT than the builder or dealer. This concludes that the Owner owns more land and sells, coming to dealer/builder, dividing the land, building more houses, and selling at the profit margin.

**Machine Learning**

Machine learning first started in 1959 by Arthur Samuel, who worked in IBM and the pioneer field of computer gaming and artificial intelligence.

Diagram

Description automatically generated

Machine learning a mixed field of statistics, computer programming, mathematics analysis. Machine learning has involved in various fields like medicine, computer vision, spam filtering, voice recognitions.

For learning the data, machine learning uses these standard practices.

* Supervised Learning
* Unsupervised Learning
* Semi-Supervised Learning
* Reinforcement Learning
* Dimensionality Reduction

**Regression**

For this project, it is best suitable for regression techniques. Regression is a method used in finance, investing. There are several regression techniques are there.

* Multiple linear regression
* Logistic Regression
* Support Vector Regression
* Random Forest Regression

In Regression, independent variables add random weights to predict the dependent variable by using the gradient descent technique to optimise weight for predicting accurately.

By applying these techniques and error-correcting methods, able to predict the prices that suitable for a house with that features.

**Artificial Intelligence**

Artificial Intelligence is machines that will try to mimic functions that humans associate with the human mind, such as "problem-solving" and "learning".

**Deep Learning**

In the early 21st century, Large storage made cheap and available to everyone. So industries utilised this opportunity as a by-product of this large amounts of data(formerly known as "big data") available now. Using this data with AI/ML companies benefitting product development, customer satisfaction, improvement in advertisements.

In deep learning, primarily neural networks were used. Neural networks are a series of algorithms that endeavours to recognise underlying relationships similar to the human mind.

## Deliverables

Every data science project follows these practices to run the project smoothly. Practising these steps were easy to understand and practise. Even this very easy to upgrade when needed.

### Practises:

Requirement:

Every project has an objective to accomplish. The data science project was not unique from that. The initial proposal was the first step of the project.

Data Collection:

The next step was collecting data from various sources and should be very reliable. Non-reliable sources of data lead to the failure of the data project.

Data Analysis:

This step is crucial to make an initial analysis of data. Here about to know that how data has spread and which independent factors influencing the dependant factors.

Data preparation:

Every data is unique and has some meaning to it. It has some gaps, and some values will go missing, or it will be a different format than the usual. These issues have addressed and fill accordingly with best practices; otherwise, modelling may not be possible.

Modelling:

Prepared data now ready to insert to machine learning models to predict. Once initial guesses are produced, the estimators and weight correction methods model will be fine-tuned to give the best predictions.

Here evaluation and comparison of other models can be made to finalise the deployment model. In this project, regression and neural networks will be implemented.

Deployment:

After performing modelling, one of the suitable models has selected, trained and deployed.

### Tools used in this project

In this project, full implementation on python using these frameworks: NumPy, pandas, matplotlib, sci-kit-learn, TensorFlow and PyTorch.

### What factors influence choosing these tools

Python is a straightforward and readable language, among others. Python language has widely adopted by many data scientists and using currently. The frameworks mentioned above was developed by other data scientists and using day-to-day life. Another important thing is open source.

## Progress

* Data is in CSV format. Now it has loaded and converted into Pandas data frame.
* Analysis was performed on the data and derived some conclusions about what factors influencing the price factor. Some conclusions have mentioned above.
* Before further processing, data must be cleaned and transformed; null checks and other format data types were converted.
* For machine learning, Categorical values are encoded and scaled.
* Then all data were analysed and made correlation and produced heatmap.
* An initial linear regression model was developed to find how it will perform.

## Problems

### Linear Regression

Linear regression performed worst on this dataset. Mean Squared Error has produced 80 error rate. This data is non-linear is to used other powerful linear regression methods like Polynomial, SVR, Random Forest regression.

## Planned Work

### Major Tasks

#### Literature review

In this step, I will research literature regarding regression and neural networks. So we can understand how other hyper-parameters tuning will boost the performance and others areas to a lookout.

Dataset Collection:

This project based on Kaggle dataset house prices predictions for India. This dataset licensed under GPL2.

#### Research different types of regression and evaluation techniques:

Here deeply research on regression and evaluation techniques.

Upon implementing this on the project, the dataset will improve predictions.

#### Research on neural networks:

While doing Machine learning and researching the neural networks can also provide a better solution for this project.

#### Conclusion:

Implementing different techniques and electing the best model to predict prices on houses based on provided input features.

#### Final Report:

The report was doing while implementing the project but the last days for finalising the formatting, indentations and type corrections.