House Market Prices Predicting With Machine Learning

Table of Contents

[Introduction 3](#_Toc76033541)

[Problem 3](#_Toc76033542)

[Hypothesis 3](#_Toc76033543)

[Background Research 3](#_Toc76033544)

[Analysis on project data 4](#_Toc76033545)

[About Machine Learning 5](#_Toc76033546)

[Regression 6](#_Toc76033547)

[Progress 6](#_Toc76033548)

[Tools used in this project 6](#_Toc76033549)

[What factors influence choosing these tools 6](#_Toc76033550)

[Project Update 6](#_Toc76033551)

[Problems 7](#_Toc76033552)

[Project Plan 7](#_Toc76033553)

[Progress up to date 7](#_Toc76033554)

[Progress further 7](#_Toc76033555)

[Addressing Legal, Ethical, Professional, Social-issues 8](#_Toc76033556)

[Legal 9](#_Toc76033557)

[Social 9](#_Toc76033558)

[Ethical 9](#_Toc76033559)

[Professional 9](#_Toc76033560)

[References / Bibliography 9](#_Toc76033561)

[Appendices 10](#_Toc76033562)

## 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.

## Background Research

### Analysis on project data

#### 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.

#### **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.

## About 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.

## Progress

### 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.

### Project Update

* 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.
* After Scaling the data, a Linear regression operation was performed. Linear regression was a basic regression technique, and it outperforms the prediction when data is linear.

### Problems

#### Linear Regression

Linear regression performed poor on this dataset. Mean Squared Error has produced 80 error rate. In conclusion, This dataset was non-linear. More accuracy prediction and less error rate need to use powerful linear regression methods like Polynomial, SVR, Random Forest regression.

## Project Plan

### Progress up to date

#### 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.

### Progress further

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

In Machine learning, There is a lot of algorithms and techniques were present. For better predictions, these practices were developed.

1. Pre-processing
2. Hyper-parameter fine-tuning
3. Estimator correction
4. Cross-validation

For this project, Due to the non-linearity of a dataset, some regression algorithms are about to use(mentioned above). Those algorithms are popularly known for complex and high dimensional datasets. Please find appendix for Gantt chart

Additional one day backup for Regression. If Some techniques will take more time so this backup day will be used. Another critical, Ensemble technique is a powerful technique combining multiple regressions in one line will increase predictions and fewer failover conditions.

|  |  |  |  |
| --- | --- | --- | --- |
| Technique | Days | From | To |
| Logistic & Polynomial | 5 days | 28/6/2021 | 4/7/2021 |
| Support Vector Regression | 5 days | 5/7/2021 | 11/7/2021 |
| Random Forest Regression | 5 days | 12/7/2021 | 18/7/2021 |
| XGBRegressor | 5 days | 19/7/2021 | 25/7/2021 |
| Back up | 1 day | 25/7/2021 | 26/7/2021 |
| Ensemble Technique | 5 days | 27/7/2021 | 1/8/2021 |

#### Research on neural networks:

Neural networks are another popular and best solution for data science projects. There are easy to implement and more powerful. For this project, these concepts of neural networks were about to implement.

|  |  |  |  |
| --- | --- | --- | --- |
| Technique | Days | From | To |
| Deep Neural Networks | 3 days | 2/8/2021 | 5/8/2021 |
| Recurrent Neural Networks | 3 days | 6/8/2021 | 9/8/2021 |
| Boltzmann machine | 3 days | 10/8/2021 | 13/8/2021 |

#### End-To-End Model

After implementing these models and finding the best-suited model/ technique. We need to put all pieces together, so the user finally uses this project.

|  |  |  |  |
| --- | --- | --- | --- |
| Technique | Days | From | To |
| Deep Neural Networks | 2 days | 14/8/2021 | 15/8/2021 |

#### Final Documentation

At every step, the final report prepared for remembering everything that progress has done. These days for formatting, type corrections and presentations will be done.

|  |  |  |  |
| --- | --- | --- | --- |
| Technique | Days | From | To |
| FPR | 10 days | 16/8/2021 | 25/8/2021 |

## Addressing Legal, Ethical, Professional, Social-issues

### Legal

This project data was collected on Kaggle, and it is licensed under GPL v2, which means fair to use and distribute according to license.

All tools and frameworks which are used in this project were open-source. In university / who is using it does not get any legal issues.

Dataset:

Kaggle.com. 2021. *House Price Prediction Challenge*. [online] Available at: <https://www.kaggle.com/anmolkumar/house-price-prediction-challenge> [Accessed 1 July 2021].

License:

Gnu.org. 2021. *GNU General Public License v2.0 - GNU Project - Free Software Foundation*. [online] Available at: <http://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html> [Accessed 1 July 2021].

### Social

No persons were involved in this project, or any other practices were made to hurt the feelings or beliefs of persons.

### Ethical

Strictly this project does not expose any gender or any other beliefs. This project aim was to help buy the best house in society according to the buyer budget.

### Professional

Under any circumstances, Maintaining confidentiality, anonymity will not be uneased.

All issues and concerns were following under the guidelines of the university. No extra practices made to disrupt the above issues.

## References / Bibliography

1. Geron, A., 2019. *Hands-On Machine Learning with Scikit-Learn, Keras and Tensor Flow: Concepts, Tools and Techniques to Build Intelligent Systems*. 2nd ed. O'REILLY, pp.186-290.
2. Subramanian, G., n.d. *Python data science cookbook*. Birmingham, Uk: packt, pp.315-376.
3. Boschetti, A. and Massaron, L., 2015. *Python data science essentials*. Birmingham: Packt, pp.123-140,205-231.
4. Medium. 2021. *Predicting House Prices with Linear Regression | Machine Learning from Scratch (Part II)*. [online] Available at: <https://towardsdatascience.com/predicting-house-prices-with-linear-regression-machine-learning-from-scratch-part-ii-47a0238aeac1> [Accessed 1 July 2021].
5. Medium. 2021. *All about Feature Scaling*. [online] Available at: <https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35> [Accessed 1 July 2021].
6. Uc-r.github.io. 2021. *Regression Artificial Neural Network · UC Business Analytics R Programming Guide*. [online] Available at: <http://uc-r.github.io/ann\_regression> [Accessed 1 July 2021].
7. Stanford.edu. 2021. *CS 230 - Recurrent Neural Networks Cheatsheet*. [online] Available at: <https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks> [Accessed 1 July 2021].

## Appendices