

NAME OF THE PROJECT:

**HOUSING: PRICE PREDICTION**

Submitted by:

Pratyush Raj

**ACKNOWLEDGMENT**

I would like to express my special gratitude to the “Flip Robo” team, who has given me this opportunity to deal with the dataset and it has helped me to improve my analysis skills. And I want to express my immense gratitude to Md. Kashif (SME Flip Robo), he is the person who has helped me to get out of all the difficulties I faced while doing the project.

A huge thanks to “Data trained” who are the reason behind my Internship at Fliprobo. Lastly, my parents and my friends have been my backbone in every step of my life.

References used in this project:

1. SCIKIT Learn Library Documentation
2. Blogs from data science, Analytics Vidya, Medium
3. Andrew Ng Notes on Machine Learning (GitHub)
4. Data Science Projects with Python Second Edition by Packt
5. Hands-on Machine learning with scikit learn and tensor flow by  
   Aurelien Geron

**Chap 1. Introduction**

**1.1 Business Problem Framing**

Real Estate Property is not only the basic need of a man but today it also represents the riches and prestige of a person. Investment in real estate generally seems to be profitable because property values do not decline rapidly. The market demand for housing is always increasing every year due to the increase in population and migration to other cities for their financial purpose. Changes in the real estate price can affect various household investors, bankers, policymakers, and many more. Investment in Housing seems to be an attractive choice for the investments.

Houses are one of the necessary needs of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors to the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help companies increase their overall revenue, and profits, improving their marketing strategies and focus on changing trends in house sales and purchases.

In general, purchasing and investing in any real estate project will involve various transactions between different parties. Thus, it could be a vital decision for both households and enterprises. How to construct a realistic model to precisely predict the price of real estate has been a challenging topic with great potential for further research.

There are many factors that have an impact on house prices, such as the number of bedrooms and bathrooms. House price depends upon its location as well. A house with great accessibility to highways, schools, malls, and employment opportunities, would have a greater price as compared to a house with no such accessibility.

Regression is a supervised learning algorithm in machine learning which is used for prediction by learning and forming a relationship between present statistical data and target value i.e., Sale Price in this case. Different factors are taken into consideration while predicting the worth of the house like location, neighborhood, and various amenities like garage space, etc. if learning is applied to the above parameters with target values for a certain geographical region as different areas differ in price like land price, housing style, the material used, and availability of public utilities.

**1.2 Conceptual Background of the Domain Problem**

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below. The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

This company wants to know:

▪ Which variables are important to predict the price of a variable?

▪ How do these variables describe the price of the house?

It is required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

**1.3 Review of Literature**

Related Work or Literature survey is the most important step in any kind of research. Before start developing Machine Learning Model, we need to study the previous papers of the domain in which we are working and on the basis of the study we can predict or generate the drawback and start working with the reference of previous papers. In this section, we briefly review the related work on house price prediction and the techniques used. Predicting house prices manually is a difficult task and generally not very accurate, hence there are many systems developed for house price prediction.

Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, and Rick Siow Mong Goh [6] proposed an advanced house prediction system using linear regression. This system’s aim was to make a model that can give us a good house price prediction based on other variables. This paper proposed on Hybrid Regression technique for housing Prices Prediction focused on the use of creative feature engineering to find the optimal features and their correlation with Sales Prices. Feature engineering improved the data normality and linearity of data. Their system showed that working on the Ames Housing dataset was convenient and showed that the use of Hybrid algorithms (65% Lasso and 35% Gradient Boost) provided results in predicting the house prices rather than using one from lasso, ridge, or gradient boost.

CH.Raga Madhuri, Anuradha G et.al [7], estimated house prices by the analysis of fare ranges, foregoing merchandise, and forewarning of developments. The author discussed diverse regression techniques such as Gradient boosting and AdaBoost Regression, Ridge, Elastic Net, Multiple linear, and LASSO to locate the most excellent. The performance measures used are [MSE] Mean Square Error and [RMSE] Root Mean Square Error.

Zhongyun, Jiang, Guoxin, and Shen [8], develop 6-layer BP neural network with Kera's deep learning. For the backend Tens flow or Theano is used. The test results give the predicted real value of the house with an accuracy of 95.59%. The Gaussian noise model is favorable for the house price forecast.

Zhen Peng, Qiang Huang, and Yincheng Han [9], precisely study the cost of recycled houses and examined 35417 bits of information caught by the Chengdu HOME LINK organization. First and foremost, the caught information was cleaned, and the attributes were chosen. The test results show that the precision of XGboost expectation is the most elevated, and the forecast exactness score arrives at 0.9251. Dissimilarity with linear regression, decision tree model, the XGboost algorithm gives improved speculation capacity and strength in information forecast, and, furthermore, data prediction over-fitting aspect, establishing a strong framework for the ensuing recycled house value expectation.

Liu et al. [10] have constructed a statistical model based on the fuzzy neural network prediction model, which incorporates the hedonic

theory and a great database with relevant characteristics affecting the price of properties based on recently sold projects. The experimental outcome and analysis have shown that the fuzzy neural network prediction model has a promising ability for real estate price prediction given reliable input data with high quality.

According to Kusan et al. [11], factors affecting housing sale prices can be classified into three types: house factors, environmental factors, and transportation factors. The most influential type is residential factors, including residence, usability, and the number of rooms. When people consider purchasing a house for living purposes, the factors above are the main determinants for living quality. Buyers with family members would typically attach more importance to the essential feature of the house, like the living area and number of rooms, which have a significant impact on the overall living quality and experience in the house. Besides, the intangible features, like the view of residence and usability, also have a rather considerable influence on the housing price, through affecting buyers’ experience on the house and willingness to pay. The other influential types are the main factors related to building properties and floor factors. Building properties are mainly about hardware and basic facilities in the building, such as the elevator, generator, and garage. The rising trend of numbers of vehicles per household possessing generates a necessary demand for the quality and capacity of a garage in a house. Other affiliated facilities to the house like the swimming pool and backyard have also played an essential role in determining the housing price, as the demand for leisure and relaxation has been arising with the economic progress. On the other hand, floor factors, like the number of floors, have also impacted the housing price significantly. A family with children and elders tends to prefer a house with multi floor construction, which offers different family members separate living areas with appropriate privacy while living together. Environmental factors mainly consist of two parts: regional environment and nearby pollution. Regional Scientific Pro environment refers to the overall living conditions in the surrounding community. Sanitation, as a significant indicator of the living quality, has been given more importance in the recent decades. A community with comprehensive sanitation services tends to attract buyers to pay a higher price.

The paper proposed by Ayush Varma et. al. [12] suggested that the use of neural networks along with linear and boosted algorithms improved prediction accuracy. Three algorithms were used namely Linear Regression, Forest Regression, and Boosted Regression. The dataset was tested on all three and the results of all the above algorithms were fed as input to the neural network. Neural networks were used mainly to compare all the predictions and display the most accurate result. A neural network along with Boosted Regression was used to increase the accuracy of the result.

Another research study showed that there exist relationships between visual appearance and non-visual attributes such as crime statistics, housing prices, population density, etc. of a city. For instance, “City Forensics: Using Visual Elements to Predict Non-Visual City Attributes” [13], uses visual attributes to predict the sale price of the property.

**1.4 Motivation for the Problem Undertaken**

The project is provided to me by Flip Robo Technologies as a part of the internship program. The exposure to real-world data and the opportunity to deploy my skillset in solving a real-time problem has been my primary motivation.

Our main objective of doing this project is to build a model to predict house prices with the help of other supporting features. In order to improve the selection of customers, the client wants some predictions that could help them in further investment and improvement in the selection of customers.

The No Free Lunch Theorem state that algorithms perform differently when they are used under the same circumstances. This study aims to analyze & predicting house prices when using Multiple linear, Lasso, Ridge, XGBoost, Random Forest regression, and Extra Tree Regressor algorithms. Thus, the purpose of this study is to deepen the knowledge in regression methods in machine learning.

In addition, the given datasets should be processed to enhance performance, which is accomplished by identifying the necessary features by applying one of the selection methods to eliminate the unwanted variables since each house has its unique features that help to estimate its price. These features may or may not be shared with all houses, which means they do not have the same influence on the house pricing resulting in inaccurate output.

The study answers the following research questions:

- Research question 1: Which machine learning algorithm performs better and has the most accurate result in house price prediction? And why?

- Research question 2: What are the factors that have affected house prices in Australia over the years?

10

Surprise Housing - Housing Price Predication & Analysis Project

**Chap 2 Analytical Problem Framing**

**1. Mathematical / Analytical Modelling of the Problem**

Our objective is to predict House prices which can be resolved by the use of the regression-based algorithm. In this project, we are going to use different types of algorithms which uses their own mathematical equation in the background. This project comes with two separate data sets for the training & testing model. Initially, data cleaning & pre-processing perform over data. Feature engineering is performed to remove unnecessary features & for dimensionality reduction. In model building, the Final model is selected based on evaluation benchmarks among different models with different algorithms. Further Hyperparameter tuning was performed to build a more accurate model out of the best model.

**2. Data Sources and their formats**

The data set provided by Flip Robo was in the format of CSV (Comma Separated Values). There are 2 data sets that are given. One is training data and one is testing data.

1) Train file will be used for training the model, i.e., the model will learn from this file. It contains all the independent variables and the target variable. The dimension of the data is 1168 rows and 81 columns.

2) Test file contains all the independent variables, but not the target variable. We will apply the model to predict the target variable for the test data. The dimension of the data is 292 rows and 80 columns.

11

Surprise Housing - Housing Price Predication & Analysis Project

The data types of different features are as shown below:

**3. Data Pre-processing**

The dataset is large and it may contain some data errors. In order to reach clean, error-free data some data cleaning & data pre-processing performed data.

• Data Integrity check –

No duplicate entries are present in the dataset.

We have removed features that contain a high amount of missing values e.g., the Top 5 features with missing values in the above list. The rest of the features are handled based on mean, median, or mode imputation depending on outliers & distribution of features.

▪ Feature extraction for age-related features.

13

Surprise Housing - Housing Price Predication & Analysis Project

▪ Some of the features are removed as they are represented in another similar feature.

▪ Label Encoding of Categorical features:

The categorical Variable in the training & testing datasets are converted into numerical datatypes using a label encoder from scikit library.

Surprise Housing - Housing Price Predication & Analysis Project ▪ Standard Scaling:

**4.Data Inputs- Logic- Output Relationships**

A correlation heatmap is plotted to gain an understanding of the relationship between target features & independent features. We can see that lot of features are highly correlated with the target variable Sale Price. To gain insights into the relationship between Input & output different types of visualization are plotted which we will see in EDA section of this report.

15**Models Development & Evaluation**

**Identification Of Possible Problem-Solving Approaches (Methods)**

Our objective is to predict house prices and analyze features impacting Sale prices. This problem can be solved using regression-based machine learning algorithms like linear regression. For that purpose, the first task is to convert categorical variables into numerical features. Once data encoding is done the data is scaled using a standard scalar. The final model is built over this scaled data. For building an ML model before implementing the regression algorithm, data is split into training & test data using train\_test\_split from the model\_selection module of sklearn library.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model. After that model is trained with various regression algorithms and 5-fold cross-validation is performed. Further Hyperparameter tuning was performed to build a more accurate model out of the best model.

**2. Testing of Identified Approaches (Algorithms)**

The different regression algorithms used in this project to build ML model is as below:

❖ Linear Regression

❖ Random Forest Regressor ❖ Decision Tree Regressor ❖ Ridge Regression

❖ XGB Regressor

❖ Extra Tree Regressor

17

Surprise Housing - Housing Price Predication & Analysis Project

**KEY METRICS FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION**

The following metrics used for evaluation:

1. Mean absolute error which gives the magnitude of difference between the prediction of observation and the true value of that observation.
2. Rootmeansquareerrorisoneofthemostcommonlyusedmeasures for evaluating the quality of predictions.
3. R2 score which tells us how accurate our model predict the result is going to important evaluation criterion along with the Cross-validation score.
4. Cross Validation Score

**RUN AND EVALUATE SELECTED MODELS**

**1. Linear Regression:**

**2. Random Forest Regressor:**

**3. Decision Tree Regressor:**

**4. Extra Trees Regressor:**

**5. XGB Regressor:**

**6. Ridge Regressor:**

5-Fold cross-validation performed over all models. We can see that Random Forest Regressor gives a maximum R2 score of 90.50 and a cross-validation score of 83.30 %. Among all models, we will select Random Forest Regressor as a final model and we will perform hyperparameter tuning over this model to enhance its R2 Score.

20

Surprise Housing - Housing Price Predication & Analysis Project Final model is built with the best params got in hyper parameter tuning.

We can see that hyperparameter tuning leads to an increase in R2 Score slightly from the default model.

**Conclusion**

**1. Key Findings and Conclusions of the Study**

* ➢ Random Forest Regressor gives us the maximum R2 Score, so Random Forest Regressor is selected as the best model.
* ➢ After hyperparameter tuning Final Model is giving us an R2 Score of 89% which is slightly improved compared to the earlier R2 score of 90%.

**2. Limitations of this work and Scope for Future**

**Work**

* ➢ ANN can be used to create a more accurate model.
* ➢ Some additional features can be added to data which enable us to  
  perform Time series analysis.