# House Price Prediction using regression

KEERTHI R
*PES1UG20CS206*
*CSE DEPARTMENT*
PES UNIVERSITY

KIRAN SS
*PES1UG20CS212*
*CSE DEPARTMENT*
PES UNIVERSITY

RAKSHUK R
*PES1UG20CS324*
*CSE DEPARTMENT*
PES UNIVERSITY

***Abstract*—In today's world for most people, investing in the right thing is a very influential decision for the expected future, doing so is a very strenuous task. This study helps us predict the house prices using multiple techniques , one of which is multiple linear regression .Multiple linear regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The dataset used in this study contain house sale prices for King County which includes Seattle. It includes homes sold between May 2014 and May 2015. We have looked at various factors affecting the prices and also analysed the features that have affected the house prices the most.**

## I. Introduction

As we know that investing plays a very important role in people's lives. One of the most common type of investing is investing in property, Real estate investment. Investing can be a very overwhelming task as investing in the right property is a very tough due to the multiple factors that contribute to determine the price for the property.

Real estate is a critical driver of economic growth worldwide. In recent years as the demand for property investment has increased, using a prediction model helps real estate investors pick a better investment. The housing market which consists of the supply and demand of houses.

Housing market includes features such as supply of housing, demand for housing, house prices , rented sector and government intervention. There are a lot of factors such as income, price of housing , cost and availability, consumer preferences, investor preferences and a lot more that help people determine the choice for investing in housing.

House price prediction is very essential and important in filling in the information gap and improve real estate efficiency. There are several approaches to determine the price of a house. One of the approach is to use multi linear regression. As there are multiple factors that affect house prices, determining what factors affect the house prices the most are also essential. Multiple linear regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable

## II. RELATED WORKS

### A. House Price Affecting Factors

There are several factors that affect house prices. In his research Rahadi,[1] et al. divide these factors into three main groups, there are physical condition, concept and location. Physical conditions are properties possessed by a house that can be observed by human senses, including the size of the house, the number of bedrooms, the availability of kitchen and garage, the availability of the garden, the area of land and buildings, and the age of the house , while the concept is an idea offered by developers who can attract potential buyers, for example, the concept of a minimalist home, healthy and green environment, and elite environment.

### B. Hedonic Pricing

Hedonic price theory[2], which assumes that the value of a property is the sum of all its attributes value . In the implementation, hedonic pricing can be implemented using regression model. Equation 1 will show the regression model in determining a price. Where, y is the predicted price, and x1, x2, xi are the attributes of a house. While a, b, ... n indicate the correlation coefficients of each variables in the determination of house prices.

### C. Swedish housing market

A study was conducted in 2015 by Nils Landberg [3]. Nils analysed the price development on the Swedish housing market and the influences of qualitative variables on Swedish house prices. Landberg has studied the impact of square meter price, population, new houses, new companies, foreign background, foreign-born, unemployment rate, the number of breaks-in, the total number of crimes, the number of available jobs ranking. According to Nils, unemployment rate, number of crimes, interest rate, and new houses have a negative effect on house prices. Landberg showed that the real estate market is not easy to be analysed compared with goods market because many alternative costs are affecting the increase in house prices. The study shows that the increase in population and qualitative variables have a positive effect on house prices. The interest rate, the average income level, GDP, and the fokus 8 In contrast, the rise in interest rates has a significant negative influence on house prices. Besides, it showed unemployment rate effects negatively on house prices, but the sale price and unemployment rate are not directly correlated with each other.

### D. A hybrid Lasso and Gradient boosting regression model

A research was conducted in 2017 by Lu, Li and Yang [4]. They examined the creative feature engineering and proposed a hybrid Lasso and Gradient boosting regression model that

Identify applicable funding agency here. If none, delete this.

promises better prediction. They used Lasso in feature selec tion. They did many iterations of feature engineering to find the optimal number of features that will improve the prediction performance. The more features they added, the better the score evaluation they receive from the website Kaggle. Hence, they added 400 features on top of the 79 given features. Furthermore, they used Lasso for feature selection to remove the unused features and found that 230 features provide the best score by running a test on Ridge, Lasso and Gradient boosting.

### *E. Comparison of Artificial neural network and multiple lin ear regression for prediction*

A study was accomplished in 2017 by Suna Akkol, Ash Akilli, Ibrahim Cemal [5], where they did a comparison of Artificial neural network and multiple linear regression for prediction. In their study, the impact of different morphological measures on live weight has been modelled by artificial neural networks and multiple linear regression analyses. They used three different back-propagation techniquesfor ANN, namely Levenberg-Marquardt, Bayesian regularisation, and Scaled conjugate. They showed that ANN is more successful than multiple linear regression in the prediction they performed

## III. PROBLEM STATEMENT

The problem statement involves difficulty in interpreting the prices as many models may give a lot of varying and inconsistent prices. With this problem we intend to improve the prediction of property prices in a way where the investor can get a clear idea of investing in a property.

### *A. DATASET*

Our dataset contains the various factors that affect the prices for housing. The data tabulation offer information of the houses that include: home id, date of home sale , price , number of bedrooms, number of bathrooms, square footage of the apartment, square footage of the land, waterfront, view, condition, grade, square footage of interior housing above and below ground level, year that the house was built, year of house's last renovation, zip code, latitude, longitude , square footage of interior housing for the nearest 15 neighbors, square footage of the land lots of the nearest 15 neighbors. It helps us visualize where most of the expensive houses are located as longitudinal and latitude is being used

### *B. Exploratory Data Analysis and visualization*

(Fig 1)As there are multiple factors that affect prices , we can see that in this graph we see a distribution in the number of bedrooms in the dataset given. We see that most of the houses containing 3 bedrooms. There are also other cases such as 7 bedrooms but those can be considered as outliers, but as there can be a possibility with luxury houses we do not ignore these cases

(Fig 2)In this graph we can see the distribution of the number of floors for the houses given in the dataset. This graph contains float values which means there might be a partial floor which is also considered.

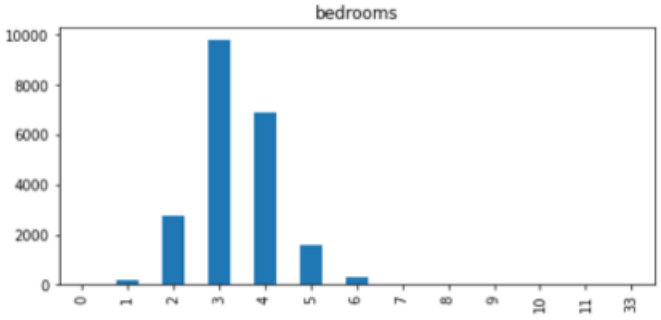


Fig. 1. Distribution of bedrooms

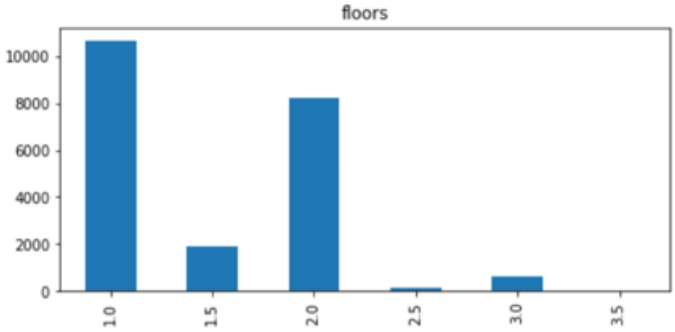


Fig. 2. Distribution of floors

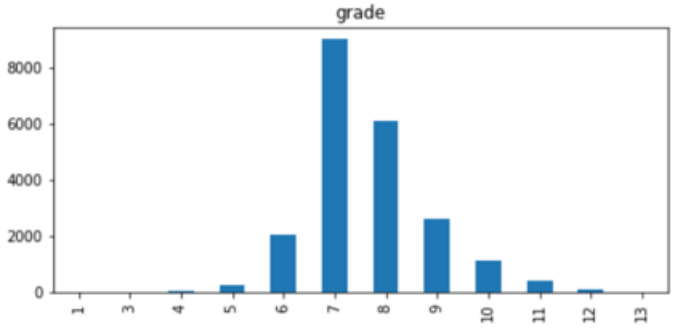


Fig. 3. Distribution of grade

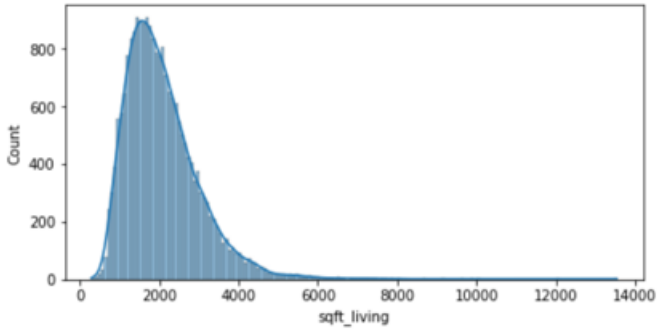


Fig. 4. Distribution plots

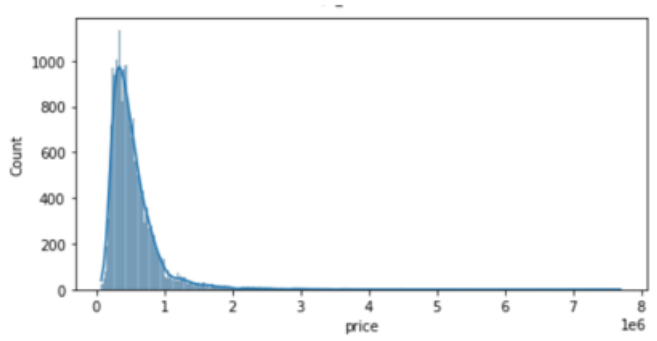


Fig. 5.

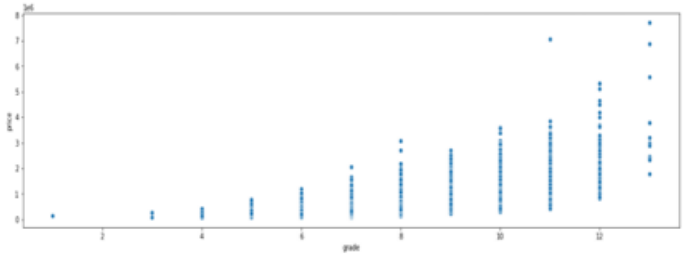

Fig. 6. Grade vs Price

(Fig 3)This graph shows the grade of the property. It can be some sort of grading given to each houses based on King County grading system, we can have a bivariate analysis for this with the price.

(Fig 4 and 5)The distribution is towards the left of the plot which can also be said as they are right skewed.

(Fig 6)As observed in the above graph , we can see that as the grade increases so does the price

(Fig 7)This map indicates the correlation of the different features of a property . As we can see from the graph that the price mainly depends on the factors such as sqftliving, sqftabove, sqftliving15, grade and bathrooms. Looking for such correlations provides very valuable information on the how the factors depend on each other.

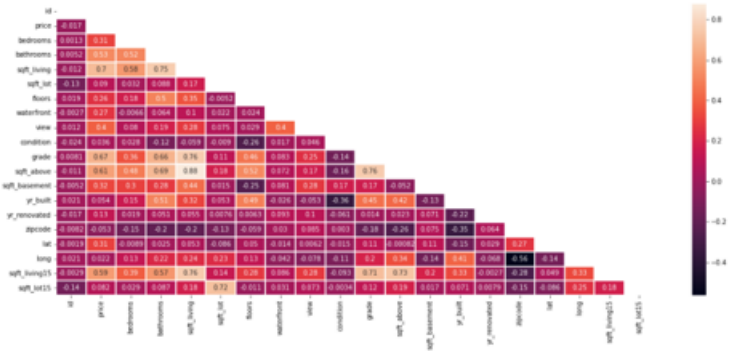

Fig. 7. Correlation Plot

## IV. PROPOSED METHODOLOGY

As the goal is to build an accurate model for house price prediction with the number of features affecting it, we have used several models to implement it . We intend to use these models to find the one which gives the most accuracy. This will help us predict the median cost of the median sized house in the area we are looking for. And finally we can rank the houses based on this data. Our model primarily aims to predict the price of houses across King County. 4 such models have been used in this project First being the linear regression model, the Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

For our prediction It works,though it does not give the best result e.g the first data is off by 100K same goes for other observation so the price varies more or less between 100K to 150K which is pretty good.

We have also displayed the actual vs predicted values , in this table we can see that it is not the best model that can be used as there is slight variation in the values.

| | Actual | Predicted |
|---|---|---|
| **735** | 365000.0 | 463830.32 |
| **2830** | 865000.0 | 772995.97 |
| **4106** | 1038000.0 | 1216910.63 |
| **16218** | 1490000.0 | 1639867.48 |
| **19964** | 711000.0 | 734666.58 |
| **1227** | 211000.0 | 265753.36 |
| **18849** | 790000.0 | 802395.12 |
| **19369** | 680000.0 | 550672.29 |
| **20164** | 384500.0 | 371955.43 |
| **7139** | 605000.0 | 473475.21 |

Fig. 8. Actual vs predicted

So, we have decided to implement the linear regression model again but this time , with selected features. Like for example selecting the index feature of the train and test data

From the prediction of this model , the predictions are decent , the result is very much close to what we have when we included all the feature to train our model, but this time the only difference is that we train it on some selected features, will say it is much better since we are getting the same result with fewer feature. (tweak the value of k estimaters to see how our model behaves).

However we have also tried other algorithms using the same features to see the difference between the previous ones. Starting with the Random Forest Regressor, Random forest is a supervised learning algorithm that uses an ensemble learning method for classification and regression. Random forest is a bagging technique and not a boosting technique. The trees in random forests run in parallel, meaning is no interaction between these trees while building the trees. And finally using the XGB regressor It is designed to be both computationally efficient (e.g. fast to execute) and highly effective, perhaps more effective than other open-source implementations. The two main reasons to use XGBoost are execution speed and model performance. XGBoost dominates structured or tabular datasets on classification and regression predictive modelling problems. The evidence is that it is the go-to algorithm for competition winners on the Kaggle competitive data science platform and this gives us the best result to work with.

## V. RESULTS AND INFERENCES

AS we implemented multiple models such as linear regression models, random forest regressor, XBG regressor, we can conclude that Random forest regressor provides the best values for the dataset.

In the figures shown we can see and compare the different Mean absolute error, mean squared error, RMSE and $R^2$ values for the models that we have implemented.(Fig. 9-12)

We have also displayed a table that compares all the models implemented and compares the RMSE and $R^2$ values.(Fig. 13)

```
Linear regression Model (including all the features)
MAE: 127896.60238400682
MSE: 44093475276.573235
RMSE: 209984.46436956528
Score (R^2): 0.6945727324117916
```

Fig. 9. Linear Regression Model(including all the features)

```
Linear regrassion model using selected features
MAE: 131205.29043204625
MSE: 46007797434.791756
RMSE: 214494.28298859566
Score (R^2): 0.6813125803734039
```

Fig. 10. Linear Regression Model(using selected features)

## VI. FUTURE PROSPECTS

The future prospect can mainly be summarized into expanding the dataset into multiple locations and areas and also including more features that affect the prices of property.As the dataset is only of a particular location, this can be expanded as there will a wider and better range of areas. We have included a a lot of different features and qualities but that too can be expanded upon. There can also be implementation of many other models on such data.

```
XGB Regressor
MAE: 88843.65777297964
MSE: 27316613030.358948
RMSE: 165277.38208950113
Score (R^2): 0.8107829236571941
```

Fig. 11. XGB Regressor

```
Random forest regressor
MAE: 89576.83036130828
MSE: 26893119352.694347
RMSE: 163991.2173035323
Score (R^2): 0.8137163852634445
```

Fig. 12. Random Forest Regressor


## ACKNOWLEDGMENT

We would like to acknowledge our Data Analytics Course Professor Anand M S for providing constant guidance during each phase of our project. We would also like to acknowledge our assistant professors who have prepared the course content and also the teaching assistants who have been constantly providing resources to practice the learnt concepts.


## REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first . . .”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published,

| | labels | RSME | R-sqr |
|---|---|---|---|
| 0 | lr | 209984.464370 | 0.694573 |
| 1 | lr1 | 214494.282989 | 0.681313 |
| 2 | rf | 163991.217304 | 0.813716 |
| 3 | xgb | 165277.382090 | 0.810783 |

Fig. 13. Comparison between all models

even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

## REFERENCES


[1] R. A. Rahadi, S. K. Wiryono, D. P. Koesrindartotoor, and I. B. Syamwil, —Factors influencing the price of housing in Indonesia,—— Int. J. Hous. Mark. Anal., vol. 8, no. 2, pp. 169–188, 2015.

[2] S. Rosen, —Hedonic Prices and Implicit Markets: Prod uct Differentiation in Pure Competition,—— J. Polit. Econ., vol. 82, no. 1, pp. 34–55, 1974.

[3] Landberg N. The Swedish Housing Market: An em pirical analysis of the price development on the Swedish housing market. Master of Science Thesis. Stockholm: KTH, Engineering and Management; 2015.

[4] Sifei Lu ZLZQXYRSMG. A Hybrid Regression Tech nique for House Prices Prediction. In 2017 IEEE International Conference on Industrial Engineering and Engineering; 2017; Singapore.

[5] Akkol S AACI. Comparison of artificial neural network and multiple linear regression for prediction of live weight in hair goats. Yyu J. Agric. Sci. 2017; 27: 21-29


## VII. PEER REVIEW

Our team was reviewed by Team SSS.

The inputs given by the team were very helpful and has helped us make changes in these ways:

1. Clarity of the problem has improved significantly via in-depth EDA.

2. Multiple MLR models and techniques have been used to ensure that the features selected produce optimal results

3. Came to a conclusion that the model random forest regressor gives us the best result and while comparing multiple linear regression models, RMSE is better choice then R square