Housing price prediction using regression

Keerthi R
PES1UG20CS206
CSE Department
PES University

Kiran S S
PES1UG20CS212
CSE Department
PES University

Rakshuk R
PES1UG20CS324
CSE Department
PES University

***Abstract*—In this part of the report we would like to show our summary of the problem statement, Exploratory Data Analysis along with different visualizations and literature review. This study helps us predict the house prices using 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 performed all the Exploratory Data Analysis that also includes few of the visualizations that we have performed to get a better understanding of the dataset.**

I. Introduction

In today's world for most people investing in housing is one of the most important decision but doing so can be a very difficult and overwhelming task due to the multiple factors that have to be taken into consideration. One of the most common type of investment that people are familiar with is real estate investment. 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 includes 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 in the Housing market. 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 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 promises better prediction. They used Lasso in feature selection. 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 linear 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 Scale conjugate. They showed that ANN is more successful tha multiple linear regression in the prediction they performed.

III. Problem Statement

A. DATASET

Our dataset contains the various factors that affect th prices for housing. The data tabulation offer information o the houses that include: home id, date of home sale , pric , number of bedrooms, number of bathrooms, square footag of the apartment, square footage of the land, waterfront, view condition, grade, square footage of interior housing above an below ground level, year that the house was built, year o house's last renovation, zip code, latitude, longitude , squar footage of interior housing for the nearest 15 neighbors, squar footage of the land lots of the nearest 15 neighbors. It help 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) This graph represents the number houses and the prices . The prices are taken in 100k to have a better visualization as seen in above graph.
- (Fig 2) This graph represents the number of bedrooms . As seen above there can be a few outliers.
- (Fig 3) The condition of the house is being shown in this respective graph. As the old houses may not be in the best condition and the newer houses will be in better condition, this graph gives a better representation.
- (Fig 4) This scatter plot represents the price by square footage. It mostly lies in the left quadrant of the graph which is where most of the area of the housing lies . The far right points beyond 4000 sqaure footage of land might be considered outliers.
- (Fig 5) This scatter plot represents how the number of bedrooms are related to the prices being at which the houses are being sold or marked at. There might be a

Fig. 1. Fig 1

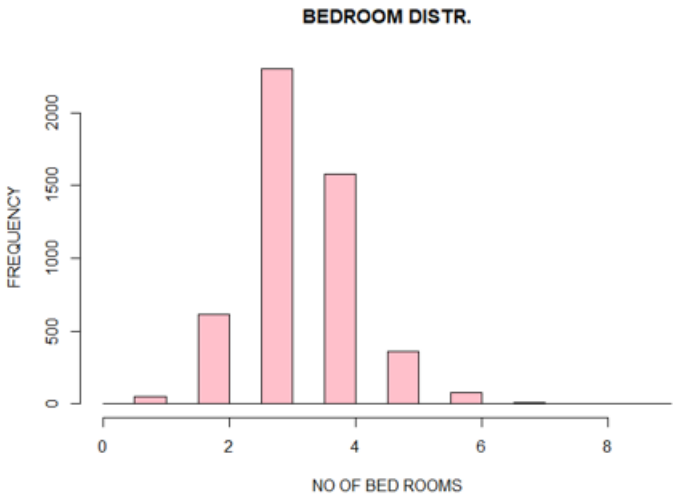

Fig. 2. Fig 2

Fig. 3. Fig 3

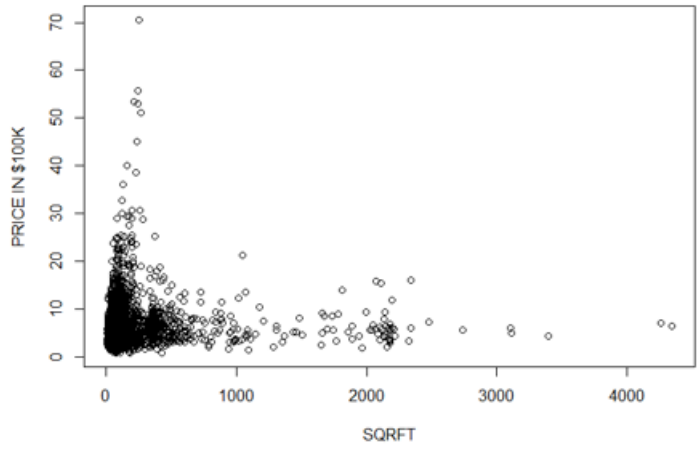

Fig. 4. Fig 4

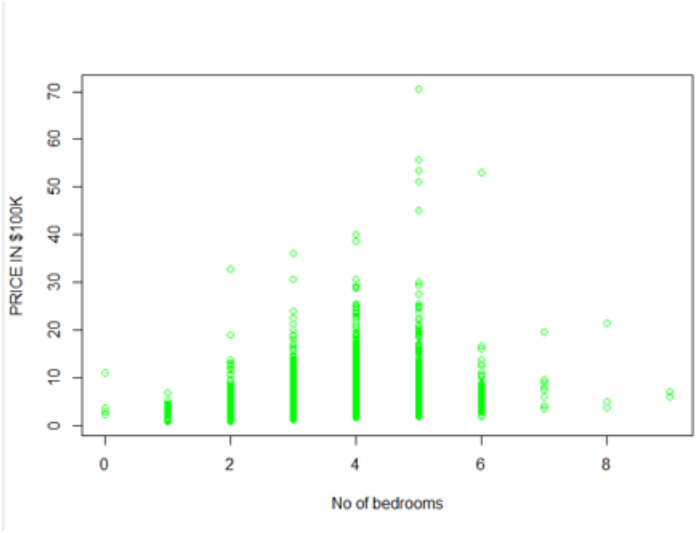

Fig. 5. Fig 5

few outliers but for the most part has an accurate representation of the relation between the number of bedrooms and how that affects the prices which is represented in terms of 100k.

IV. Proposed Methodology

As there are multiple ways to build a model for predicting housing prices, we intend to select an accurate model for predicting house prices by using multiple linear regression model. With the help of this model we can help predict the housing prices with how the various factors affect the pricing of a house such as the number of bedrooms, depending on the square footage of the property, the condition of the house and many other factors. It is mainly based in one area but considers various different factors and helps us visualize the how and where the most expensive houses lie and how thew factors affect them.

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

[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: Product Differentiation in Pure Competition,‖ J. Polit. Econ., vol. 82, no. 1, pp. 34–55, 1974.

[3] Landberg N. The Swedish Housing Market: An empirical 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 Technique 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.