

Predictive analytics for analysing the house sales (housing market)

[Business](#), [Industries](#)



Abstract: The Real Estate business is one of the rapidly growing markets in the United States of America. This is due to the growing demand for the residential sector every year, mostly in the developed Regions. Forecasting house sales is very essential for both buyers and dealers. Even it is also useful for a banker in estimating the property value while giving house loans. This research focused especially on the housing market of King County, USA. In this project, the regression analysis was performed to predict the house sales. It is one of the basic statistical methods which will be used to estimate the relationships among the different variables effect on the price of the house.

Introduction: Among all sectors, the residential sector is strongly associated with the economic development of the United States of America. One-fourth of commercial properties in America endured by the non-financial private sector are in the housing market. The demand for house marketing has a continuous increase due to this the prices of houses are rising rapidly. House is one of the basic need of the Human being's related to food, shelter and clothing. Before buying any residential property, investors have certain amount of views on various aspects of the house in terms of property type, price, location etc.

There are number of factors to estimate the price of the house such as property type, location, number of bedrooms, number of bathrooms, area of the land and other amusing facilities related to house. Both real estate agents and customers will get numerous advantages from predictive house sales. Technological Innovations have made the marketing as a reality in

recent years. To enhance customer relationship management, various data mining algorithms are used in various sectors. To find the hidden patterns in the historic data and making effective decisions for the future sales is one of the prominent benefit of the data mining techniques. In this project, regression methods are used to forecast the price index of the houses. Linear regression model was employed to find the relationship between the variables effecting the sale of the properties. Not only Regression models but also several other techniques like Neural Networks, Markov Chain and other classification methods can be used to estimate the house prices.

Related Work (Literature Review): According to the research undertaken by Hasan, Artificial Neural Networks (ANN) and hedonic regression methods were used to analyse the sale of the house by using location attribute. The results proved that the accuracy of ANN was far better than the hedonic algorithm to predict the sale of the house. The research project done by Ruth, used K-means clustering technique to extract the fuzzy patterns hidden in the residential data. Using data driven approach the outcome of the work produced better results to predict the house marketing.

Background: The dataset used in the project was gathered from the Kaggle competition. It mainly consists of house sales information of King county, USA collected in between may 2014 and may 2015. The dataset clearly provides the essential attributes which shows effect on the price of the property. Each property has a unique Id and other details like number of bedrooms, price, bathrooms, year-built, square-foot living etc. And other attributes like waterfront, view, condition, grade are also described in the

dataset. In this process, to predict the sale of the house the factors used are price, bedrooms and bathrooms and location of the house. Totally it contains twenty one columns and provided with 21631 observations with 19 types of house features.

Data processing:

- The dataset was collected from the Kaggle site.
- The downloaded data has been imported to excel and required cleaning was performed by using excel functions like trim, proper to remove messy data.
- The cleaned data has been imported to R Studio and categorical variables has been encoded to get deep insights of the dataset.
- The data has been uploaded to SPSS software for analysing using linear regression.
- In the final step, to do further analysis of data it has been coded in R.

Methodology: To predict the sale of the house it is better to review the all fields related to price of the property. Simple Regression techniques has been employed to forecast the house marketing. In this models, correlation has been computed among the attributes and shows the percentage of contribution of each attribute effecting the price of the house.

Linear Regression: Linear Regression is a basic statistical method used to find the relationship among the attributes. This technique commonly used for predictive analysis. It produces the relationship between the predictor variable and outcome variable. Through this methodology, the strength of the independent variable is determined by measuring how it is predicting the

dependent variable. And it shows the individual attribute contribution in analysing the dependent variable. The variable which is predicting is called as dependent and the variable which is using to predict is called Independent variable. Simple, it predicts the trends and shows the effects for future sales. In this process, the price of the house is considered as dependent and remaining other related attributes comes under Independent variables. This project aims to forecast the sales of the house in terms better accuracy. The linear regression is computed by using following formula: $Y = a + bX$ Where, Y = predicted outcome variable A = intercept B = slope of the line X = predictor variables In this formulate the value of Y will be calculated by using different predictor variables. From this the contribution of each predictor should be measured to predict the dependent variable. In implementation stage of the process this equation was used to calculate the required values.