House Price Prediction Using Linear Regression

Varun Goel[1], Aayna Aggarwal[2], Rupinder Kaur[3]
Department of Information Technology
[1] Assistant Professor, Maharaja Agrasen Institute Of Technology Delhi,India
[2,3] Student, Maharaja Agrasen Institute Of Technology Delhi,India

Abstract- *An accurate prediction on the house price is important to prospective homeowners, developers, investors, appraisers, tax assessors and other real estate market participants, such as, mortgage lenders and insurers . Therefore, the availability of a house price prediction model helps fill up an important information gap and improve the efficiency of the real estate market. The objective of this project to develop a basic linear regression model where a buyer inputs the relevant details about the features of the house and is provided with a predicted price on the basis of previous results*

I. INTRODUCTION

Traditional house price prediction is based on cost and sale price comparison lacking of an accepted standard and a certification process.Big data analytics is a complex process of examining large and varied data sets to uncover information including hidden patterns, unknown correlations, market trends and customer preferences that can help organizations make informed business decisions. It is a form of advanced analytics which involves a combination of predictive models, statistical algorithms and what-if analysis powered by high-performance analytics systems and techniques to analyse data sets and draw conclusions to help organizations make informed business decisions. Big data analytics involve the joint and coordinated operation of data scientists, predictive modellers, statisticians and other analytics professionals to analyse growing volumes of data.

II. EASE OF USE

Methodlogy

- Identifying the primary parameter effecting prices
- Splitting data into training set and test set
- Linear Regression technique
- Model development
- Calculating precision of model
- Testing of the model on various test data

Scope

The The data sets analysed under big data analytics are a mix of structured transaction data, semistructured and unstructured data, such as internet click stream data, web server logs, social media content, text from customer emails and survey responses, mobile phone records, and machine data captured by sensors connected to the internet of things, plus other forms of data that are often left untapped by conventional business intelligence (BI) and analytics programs. These are then organized, configured and partitioned properly. Once the data is ready, it can be analyzed with specialised softwares commonly used for advanced analytics processes. This includes tools for:

- data mining, which sift through data sets in search of patterns and relationships;
- predictive analytics, which build models to forecast customer behavior and other future developments;
- machine learning, which taps algorithms to analyze large data sets; and
- deep learning, a more advanced offshoot of machine learning.

Text mining and statistical analysis software can also play a role in the big data analytics process, as can mainstream BI software and data visualization tools.

III. OBJECTIVE

The objective of this project is to objective to develop a basic linear regression model where a buyer inputs the relevant details about the features of the house and is provided with a predicted price on the basis of previous results .Linear regression is the most basic type of regression and commonly used predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome variable? Is the model using the predictors accounting for the variability in the changes in the dependent variable? (2) Which variables in particular are significant predictors of the dependent variable?:

Abbreviations and Acronyms
,BEDROOMABVGR,BSMTEXPOSURE,BSMTFINTYPE1, BSMTQUAL,EXTERIOR1ST,EXTERQUAL,FIREPLACEQU,GARAGECARS,GRLIVAREA,HOUSESTYLE,KITCHENQUAL,LOTAREA,MASVNRTYPE,MSSUBCLASS,MSZONING,VERALLCOND,TOTRMSABVGRD,YEARBUILT
are the variables that positively and significantly affect the Sale Price, so that a rise in these variables will also increase the sales price.

RMSE

On the basis of this model, a within sample estimate is done, alternatively with 750 and 1000 observations. It was then found that the sub sample of 1000 observations yields the lowest Root Mean Square Error (RMSE) and also the lowest Thiel Inequality Coefficient and Bias Proportion.

Equations

To show how regression algorithm works we'll take into account various parameters to predict sales price. It's logical to suppose that there is a linear relationship between the parameters and price. And as we know, a linear relationship is represented by a linear equation:
y = k0 + k1*x1+k2*x2+k3*x3…..kn*xn

In our case, y equals sales price and x1,x2,x3…xn equals the various parameters. Predicting the price of a home is as simple as solving the equation (where k0,k1,k2,k3…kn are constant coefficients)

Calculations

We can calculate these coefficients using regression. Let's assume we have 1000 known house prices in a given area. Using a learning technique, we can find a set of coefficient values. Once found, we can plug in different area values to predict the resulting price.

IV. DATA DESCRIPTION AND CONCLUSIONS

The data used in the present study is obtained from www.kaggle.com. It is a data on housing prices in the state of Iowa, USA, for almost 1400 houses, based on 79 explanatory variables describing almost every aspect of residential homes in Ames, Iowa.

The Dependent Variable

Sale Price: This forms the dependent variable. Drawing a histogram reveals that it is a positive skewed distribution with a long tail to the right and with the mean>median. This implies that maximum number of houses are sold towards the lower end of the price range. The highest frequency, the mode, of the distribution is obtained in the 120,000-140,000 range in the overall price range of [34,900 ; 755,000].

The Independent Variable

Eighty one variable are taken to calculate the sale price intially. Seeing the effective probability its been filtered to 49 major variables.

Figures and Tables

Dependent Variable: SALEPRICE
Method: Least Squares

Sample: 1 1000
Included observations: 994

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	473535.6	127695.9	3.708308	0.0002
BSMTEXPOSURE	1853.353	237.6543	7.798529	0.0000
BSMTFINTYPE1	214.7179	59.30245	3.620726	0.0003
BSMTQUAL	1614.719	351.6271	4.592134	0.0000
EXTERQUAL	4733.905	600.5527	7.882580	0.0000
FIREPLACEQU	351.9304	72.11940	4.879831	0.0000
GARAGECARS	11247.65	2039.733	5.514277	0.0000
GRLIVAREA	68.17785	2.912872	23.40572	0.0000
HOUSESTYLE	815.2475	217.3485	3.750877	0.0002
KITCHENQUAL	1298.951	252.0969	5.152585	0.0000
LOTAREA	0.465568	0.102218	4.554658	0.0000
MASVNRTYPE	721.1049	228.1271	3.160978	0.0016
MSSUBCLASS	161.3944	29.22713	5.522074	0.0000
MSZONING	303.2786	147.1520	2.060988	0.0396
OVERALLCOND	4508.526	1114.586	4.045023	0.0001
YEARBUILT	194.5400	64.59078	3.011885	0.0027

R-squared	0.826477	Mean dependent var	182187.7
Adjusted R-squared	0.823816	S.D. dependent var	80518.45
S.E. of regression	33797.05	Akaike info criterion	23.71010
Sum squared resid	1.12E+12	Schwarz criterion	23.78900
Log likelihood	-11767.92	Hannan-Quinn criter.	23.74010
F-statistic	310.5434	Durbin-Watson stat	2.037892
Prob(F-statistic)	0.000000		

Table1: Coefficients of various variables

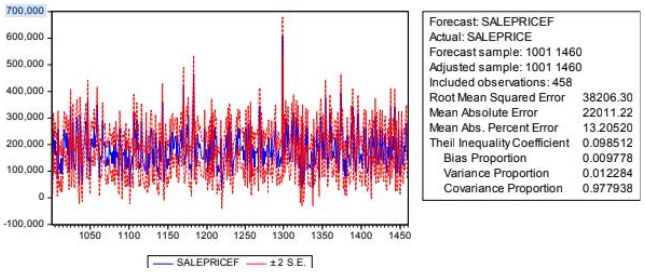

Fig: SalePrice Prediction Scatterplot

REFERENCES

[1] DONG NGUYEN, NOAH A. SMITH, CAROLYN P. ROŚE. AUTHOR AGE PREDICTION FROM TEXT USING LINEAR REGRESSION. LANGUAGE TECHNOLOGIES INSTITUTE, CARNEGIE MELLON UNIVERSITY, PITTSBURGH, PA 15213, USASEN, JAYDIP. (2018).

[2] STOCK PRICE PREDICTION USING MACHINE LEARNING AND DEEP LEARNING FRAMEWORKS.KANYONGO, G. Y., CERTO, J., & LAUNCELOT, B. L. (2006).

[3] USING REGRESSION ANALYSIS TO ESTABLISH THE RELATIONSHIP BETWEEN HOME ENVIRONMENT AND READING ACHIEVEMENT: A CASE OF ZIMBABWE. INTERNATIONAL EDUCATIONAL JOURNAL, 7(5), 632–641.