Regression Analysis for House Prices

Rohan Sinha (230041030)
Maanas (230041020)
L Praveen Kumar (230041018)

Table of Contents

What is Regression Analysis?

- A powerful statistical tool used to model the relationship between a dependent variable and one or more independent variables.
- Helps in prediction and forecasting.
- In our context, it is used to predict housing prices based on measurable attributes.

Project Objective

- Analyze how features like square footage, number of bedrooms, location, and age affect house prices.
- Build a predictive model using multiple linear regression.
- Evaluate the accuracy of the model using residual analysis.

Dataset Description

- Total records: 20 houses.
- Columns:
  - Square Footage (in sq. ft.)
  - Bedrooms (count)
  - Location (Urban/Rural)
  - Age (years)
  - Price (in dollars)
- Sourced from a Kaggle dataset.

Dataset (Raw)

SqFt	Bedrooms	Location	Age	Price
2000	3	Urban	5	450
1500	2	Urban	10	300
1800	3	Urban	8	400
1200	2	Rural	15	180
2200	4	Urban	3	420
1600	3	Urban	12	350
1400	2	Rural	20	170
2500	4	Urban	2	500
1700	3	Urban	7	360
1300	2	Rural	18	160
1900	3	Urban	6	380
2100	4	Urban	4	460
1550	3	Urban	11	320
1350	2	Rural	16	175
2300	4	Urban	1	490
1450	2	Urban	9	290
1750	3	Rural	14	240
2400	4	Urban	3	480
1600	2	Urban	10	310
1250	2	Rural	22	150

Binary Encoding of Location

To use the categorical variable `Location`, we applied one-hot encoding:

- `Urban` $\Rightarrow$ IsUrban $= 1$, IsRural $= 0$
- `Rural` $\Rightarrow$ IsUrban $= 0$, IsRural $= 1$

Note: IsRural was dropped due to multicollinearity.

Feature Matrix After Encoding

SqFt	Bedrooms	IsUrban	Age	Price
2000	3	1	5	450
1500	2	1	10	300
1800	3	1	8	400
1200	2	0	15	180
2200	4	1	3	420
1600	3	1	12	350
1400	2	0	20	170
2500	4	1	2	500
1700	3	1	7	360
1300	2	0	18	160
1900	3	1	6	380
2100	4	1	4	460
1550	3	1	11	320
1350	2	0	16	175
2300	4	1	1	490
1450	2	1	9	290
1750	3	0	14	240
2400	4	1	3	480
1600	2	1	10	310
1250	2	0	22	150

Regression Model Overview

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$

- x_1: Square Footage
- x_2: Bedrooms
- x_3: IsUrban
- x_4: Age

Estimated Coefficients

$$\beta = [28.23,\ 0.12,\ 21.79,\ 91.47,\ -3.51]$$

- Intercept: 28.23
- SqFt: 0.12 per square foot
- Bedrooms: +21.79 per room
- Urban: +91.47 for urban homes
- Age: -3.51 per year of age

Final Regression Equation

$$\hat{y} = 28.23 + 0.12x_1 + 21.79x_2 + 91.47x_3 - 3.51x_4$$

Residual Analysis

- Residual Sum of Squares (SSR): 6969.78
- Degrees of Freedom: 15 (n - p)
- Variance Estimate: $\hat{\sigma}^2 = \frac{SSR}{15} = 464.65$
- Standard Error: 20.06

Model Fit Statistics

- Coefficient of Determination (R^2): 0.974
- Adjusted R^2: 0.967
- Indicates that approximately 97.4% of the variability in house prices is explained by the model.

Residual Table

Actual (y)	Predicted ($\hat{y}$)	Residual (y - $\hat{y}$)
450	407.73	42.27
300	308.34	-8.34
400	373.18	26.82
180	163.29	16.71
420	460.56	-40.56
350	335.13	14.87
170	169.76	0.24
500	500.10	-0.10
360	364.68	-4.68
160	164.77	-4.77
380	392.21	-12.21
460	445.04	14.96
320	332.63	-12.63
175	177.80	-2.80
490	479.59	10.41
290	305.85	-15.85
240	254.65	-14.65
480	484.58	-4.58
310	320.35	-10.35
150	144.73	5.27

Visualization

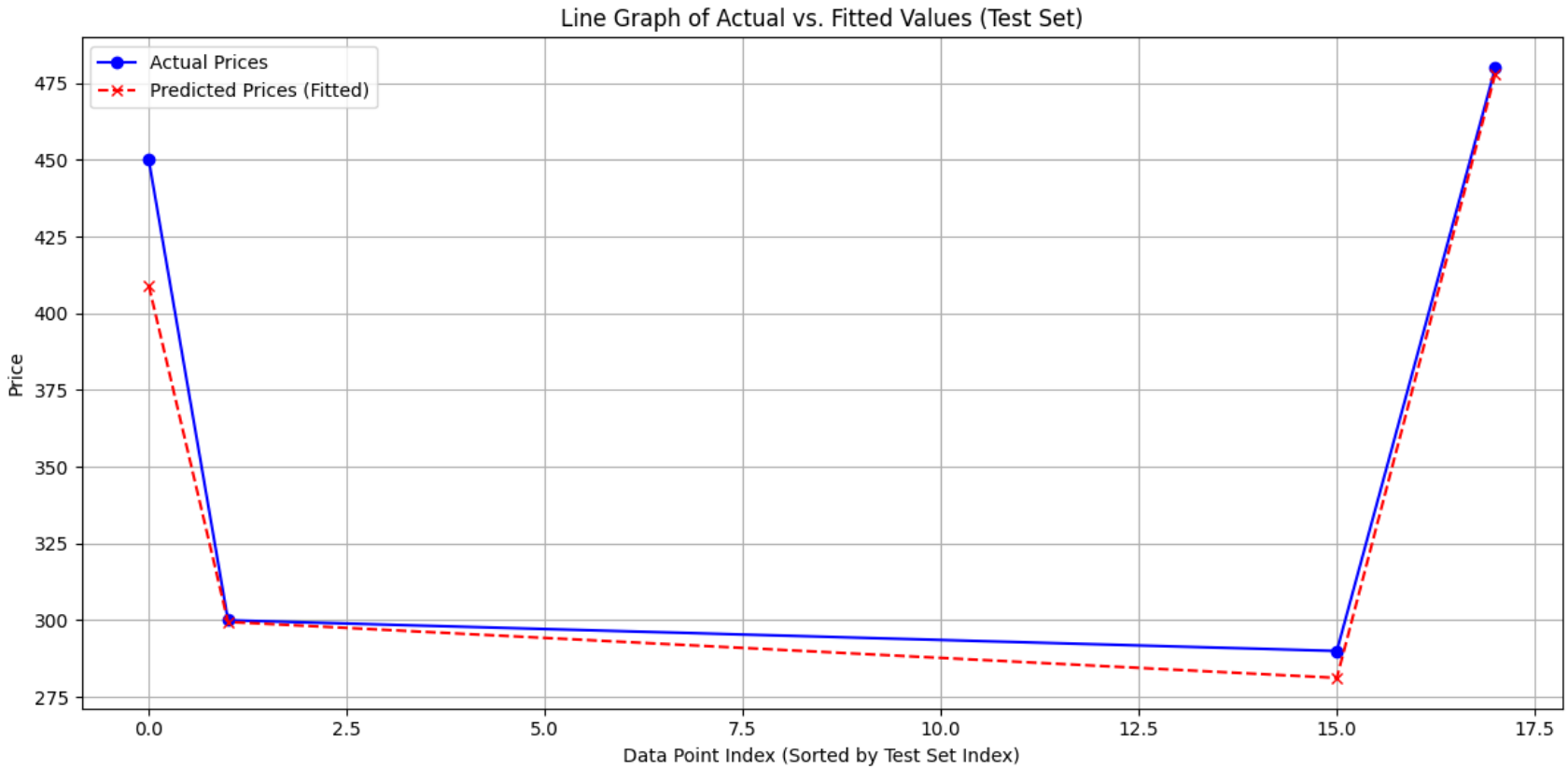

Figure: Actual vs Predicted House Prices

Visualization

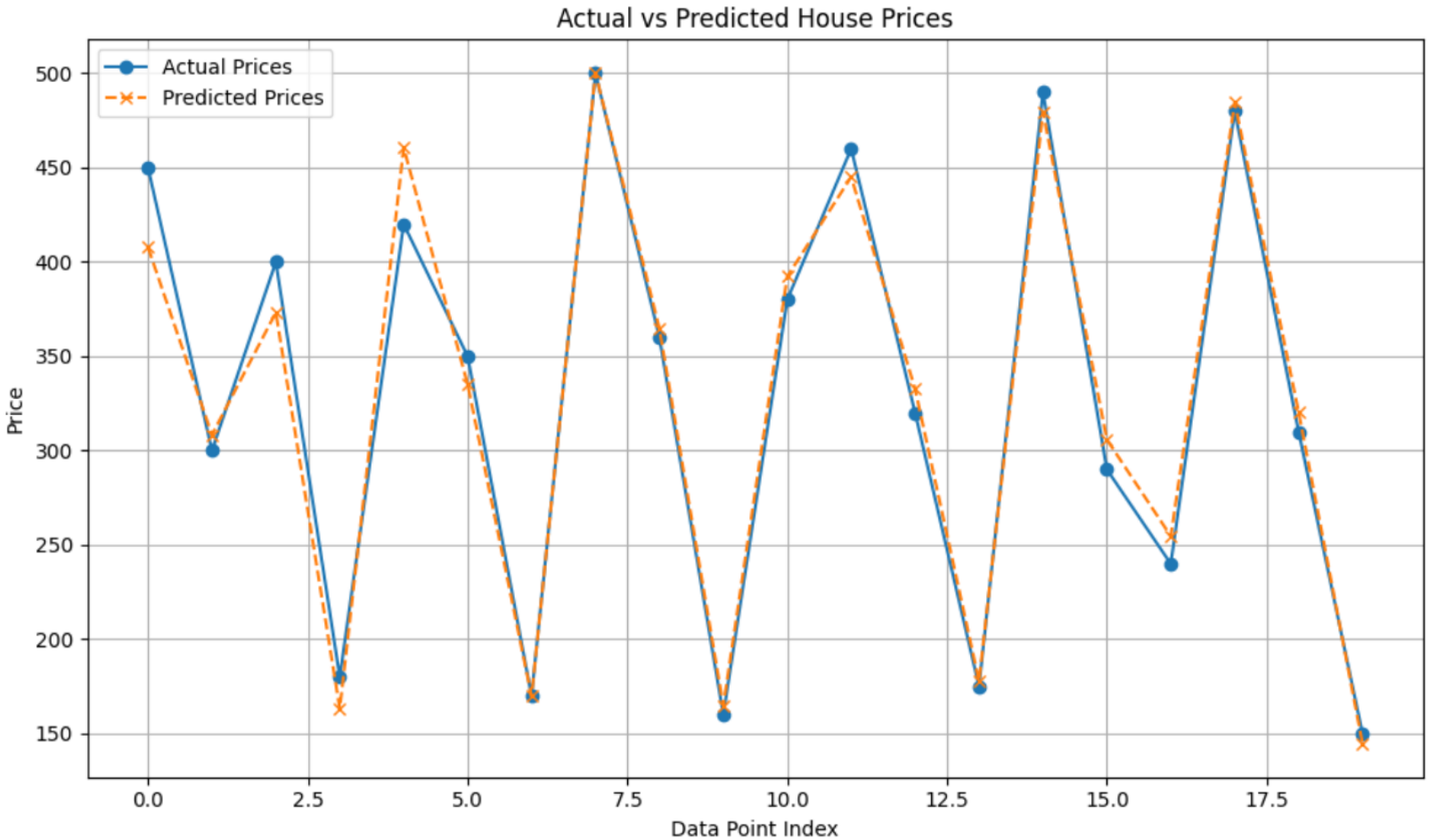

Figure: Actual vs Predicted House Prices

Conclusion

- The regression model fits the data with reasonable accuracy.
- Urban location and number of bedrooms have the highest positive influence.
- Older properties tend to have lower market value.
- Can be extended to include more features (e.g., amenities, locality rating).

References

1. Montgomery, D.C., Peck, E.A., Vining, G.G. (2012). *Introduction to Linear Regression Analysis*.
2. Kaggle Dataset: `https://www.kaggle.com/datasets/zafarali27/house-price-prediction-dataset`