Lab 1: Predicting House Prices Using Regression Model

Abhishek Sharma

DeVos Graduate School, Northwood University

MGT 665: Solving Bus Problems W/ Machine Learning

Dr. Itauma Itauma

06/08/2025

Abstract

Accurately forecasting residential property prices is important to help investors, real estate professionals, homeowners and market analysts make informed choices. This research correctly demonstrates the use of regression methods, Linear Regression, Ridge Regression, and Polynomial Regression, to predict residential property prices using a comprehensive housing dataset from Kaggle. Data processing, including parsing and normalizing different area units, square feet and square meters, cleaning currency values in lakh and crore, removing inconsistent floor row data, encoding categorical variables using the one hot method, and creating imputed values for missing values by replacing them with median or logical defaults, was extensive. The exploratory data analysis (EDA) demonstrated that there were meaningful relationships, indicating bathroom number or number of super area are significant predictors of property prices. The models were evaluated using the Mean Squared Error (MSE) metric and the Root Mean Squared Error (RMSE) metric. Ridge Regression was nearly identical in performance and showed that regularization had little impact, while the Polynomial regression was severely overfit. This exhibit provides good evidence for use of simple linear structures for predicting property prices, although it can still be further developed on model accuracy through target transformation, hyperparameter tuning, and more robust ensemble methods. This also shows the practical significance of careful data pre-processing and feature engineering in real estate analytics.

Introduction

Estimates of residential property values are important for investors, property portfolio analysis, and strategic planning in the housing market. Traditional approaches to property valuation commonly depend on market comparable approaches, known as comps, which depend on subjective human judgments that can be inconsistent. Regression and machine learning systems provide objective data methods to estimate property prices. This research

takes advantage of a housing dataset (Juhibhojani, 2025) to examine how regression systems can predict property values using the information contained in it and describe the processes of data preparation, exploratory analysis, and model evaluation.

Methodology

Data Pre-processing and Feature Engineering

The following features are mentioned in the dataset that has 187,531 property listings- Carpet Area, Super Area, Property Status, Floor details, Transaction Type, Furnishing, Facing, Overlooking, Society Name, Bathroom and Balcony counts, Ownership Type, Price information (Sale Value and Rent).

To ensure a solid modelling process, pre-processing and feature engineering was done. Parsing the area attributes, Initially, the Carpet and Super areas were recorded in different units (i.e., sqft and sqm) and formatted differently (i.e., commas, different units). Everything was converted to square feet, using one standard unit of measure (1 sqm = 10.7639 sqft), and numeric cleaning of text was done using regex and pandas.

Sale Price Conversion, the dependent variable, "Amount (in rupees)", had some currency notation issues and were not entirely consistent (Lac, Cr). These values were simply made into absolute rupee values, so that it could be used for model training (1 Lac = 1×10^5 ₹, 1 Cr = 1×10^7 ₹).

Extracting Floor Information, the original Floor data had multiple formats (as Floor information can vary in formats), there were entries for "Ground," numeric, and combination entries like "3 out of 22". Basements were omitted when possible, "Ground" floors were given numeric zero, and the floor number was extracted for proper analysis.

Categorical Variable Encoding included One-hot encoding which was utilized for categorical variables with low-cardinalities (Transaction Type, Furnishing, Facing,

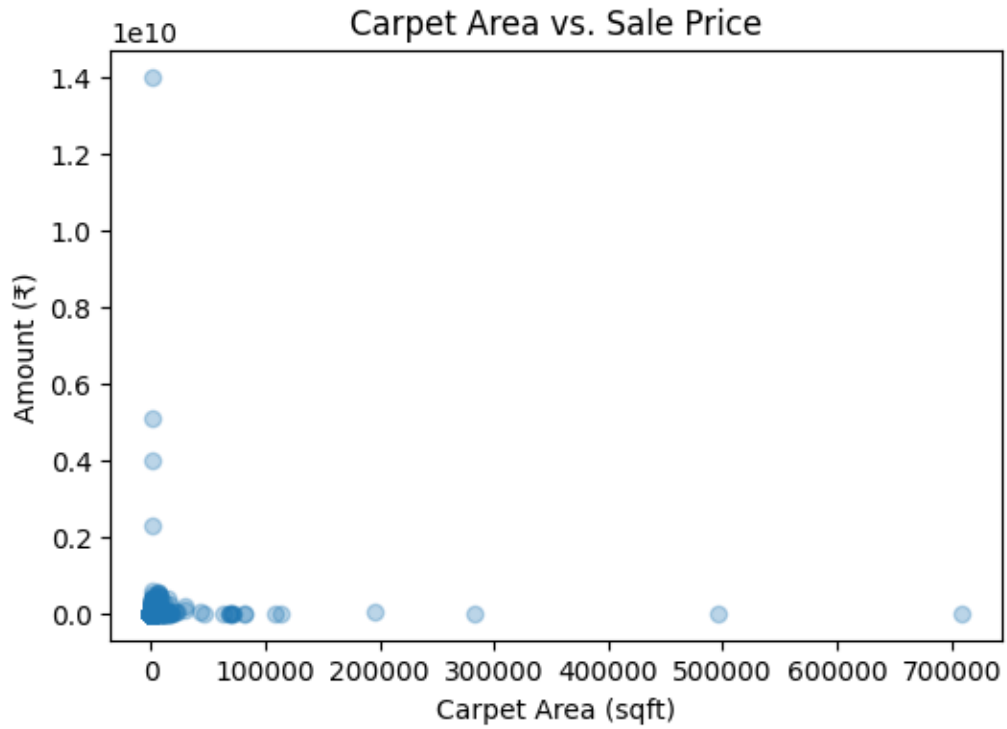
Overlooking, ownership, Property Status). For attributes with high-cardinalities (i.e., Society Name, Location Description, Title, Dimensions, Plot Area), omitted these attributes altogether to lessen potential noise and complexity in our computation.

Missing Values, the original numeric values (Bathroom, Balcony, Carpet Area, Super Area, Floor Number) with missing values were imputed with the correct medians or defaults (count of Balcony = zero). Any listing with no sale price was omitted, resulting in 177,512 observations for analysis.

Exploratory Data Analysis

Exploratory Data Analysis was done thoroughly so that relationships can be seen between the house prices, the target variable and various predictor variables.

Figure 1: *Scatter plot of carpet area vs sale price*

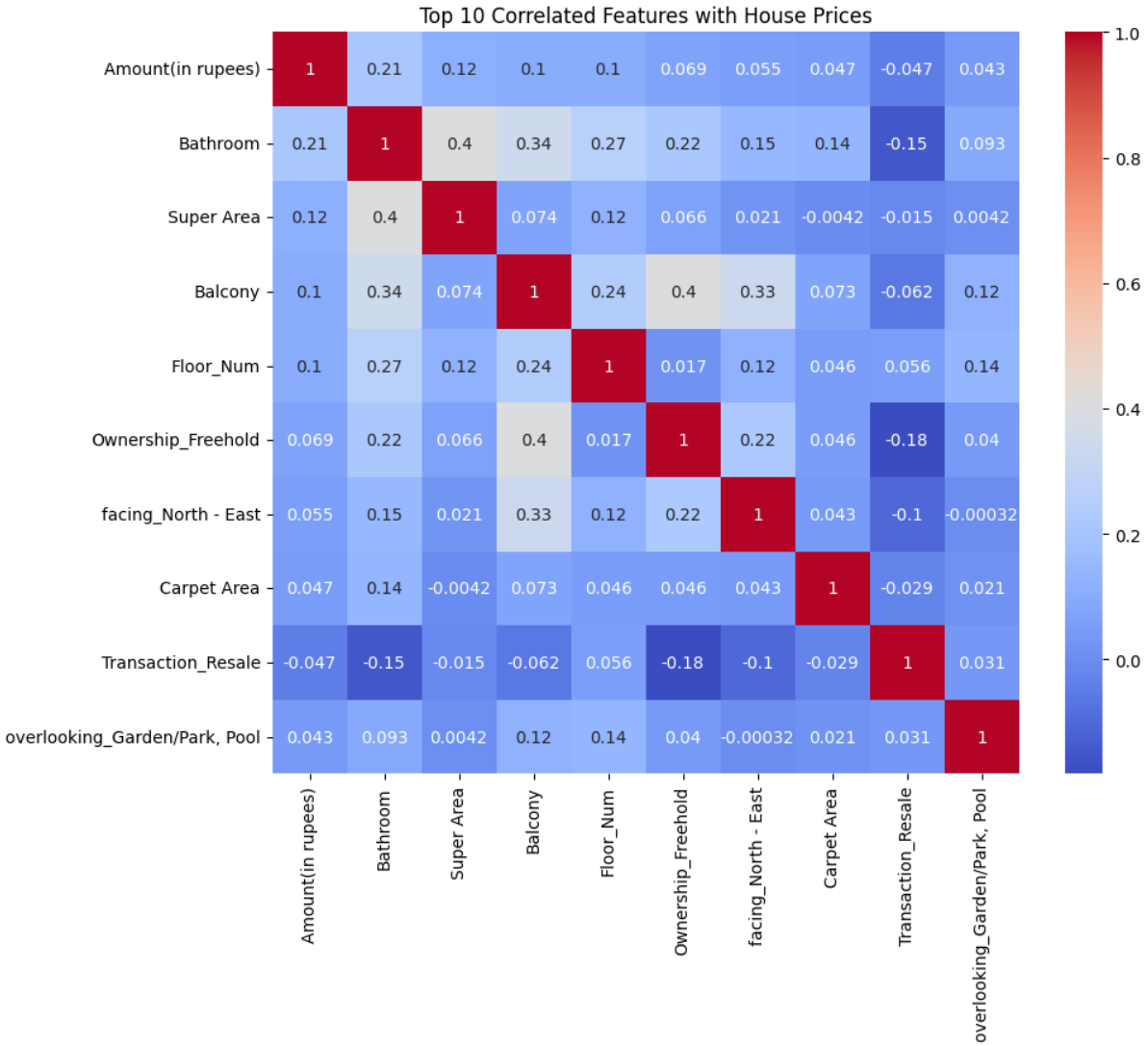


Note: The scatter plot of Carpet Area and Sale Price shows a clear indication of outliers in the data, based on the wide spread of the data points (see Figure 1).

It is clear that there were many properties at a lower carpet area with most properties less than 50,000 square feet and priced under ₹50 million. However, a few data points

represented the very largest properties (greater than 100,000 square feet) that were very high sale prices (greater than ₹1 billion) as outliers. These outliers present concern for data entry errors and/or representing the very rare property types (i.e., luxury estates or commercial real estate), which indicates the need for further outlier detection and data removal in future analyses.

Figure 2: *Correlation Heatmap with House Prices*

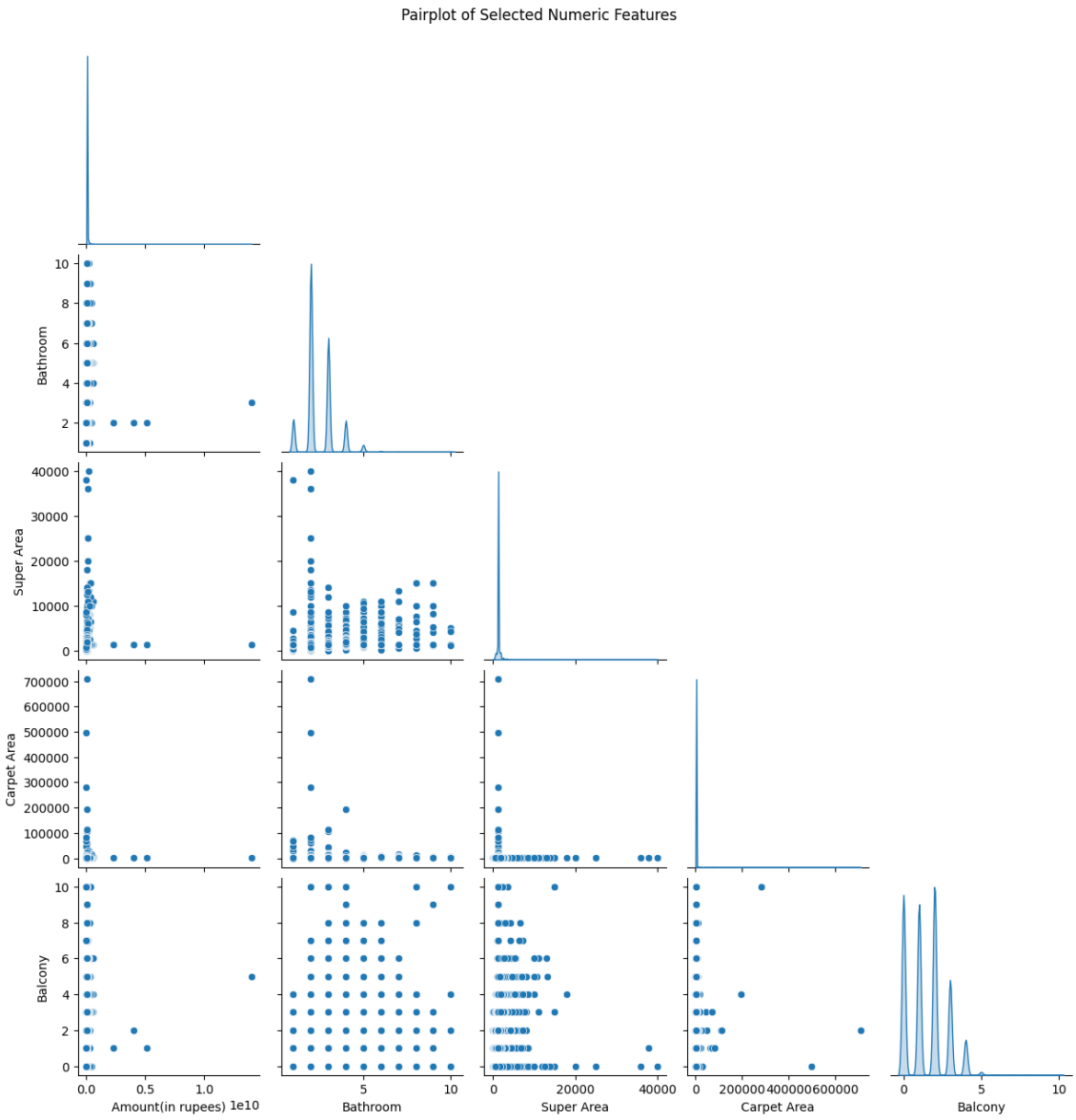


Note: A correlation heatmap was produced with the intent to identify the ten highest-level features correlated with the final house price.

The features that had the relatively stronger correlations are outline as Bathroom Count (r = 0.21), The strongest positive correlation suggested that having a greater number of bathrooms generally indicates greater market value for the property. Super Area (r = 0.12): A moderately positive correlation pointed to the observation that larger total area generally would build higher prices. Balcony (r = 0.10), A smaller yet considerable positive correlation noted that having additional balconies would build slightly higher property valuation. Floor Number (r = 0.10), Higher floors had some modest positive effect on property price with indication that some people preferred living in higher floor apartments.

Ownership type, facing direction, and resale transactions had relatively low correlations indicating that although those features will influence buyer decision, they do not necessarily play a major role in driving price variation.

Figure 3: *Pairplot of selected numeric features*

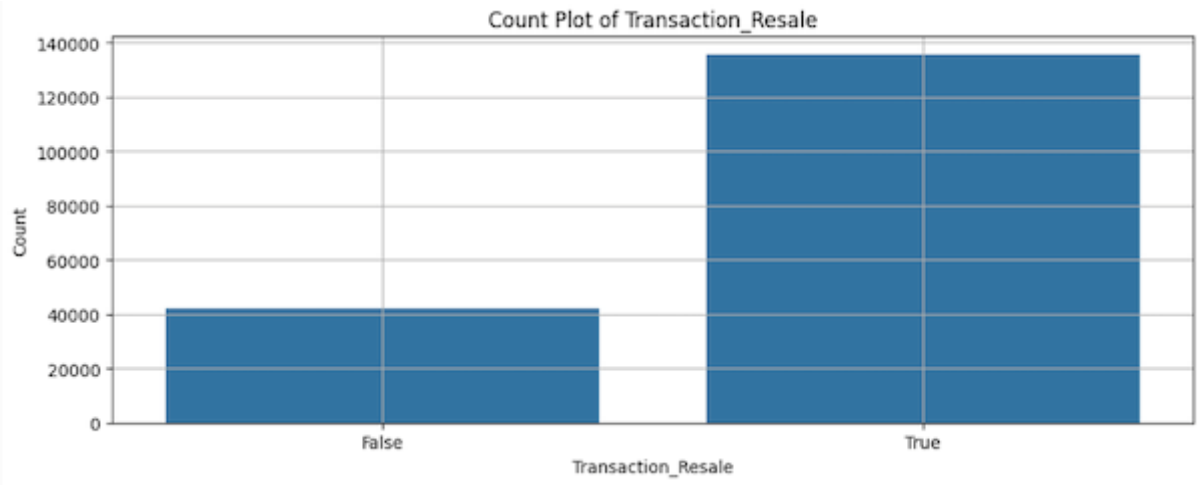


Note: The pair plot provides a visual summary of the pairwise relationships between the key numeric variables used in this analysis, Sale Price, count of Bathroom, Super Area, Carpet Area, and count of Balcony.

The following are some observations from the pair plot, Sale Price vs. Bathroom: There is a clear upward trend suggesting that as the number of bathrooms increases the price also increases. Sale Price vs. Super Area, Although a positive relationship exists there is substantial variation indicating that the super area has an impact on price, but it does not

linearly impact the price of a property, which may be due to location or luxury features. Sale Price vs. Carpet Area, There is a large amount of variance suggesting that one could not accurately predict price if only looking at carpet area and would need to consider other features.

Figure 4: *Count Plot of transaction_resale*



Note: A count plot of the Transaction_Resale feature (Figures 4) clearly shows a focus on resale transactions and new builds. There are just shy of 140,000 properties that are resale transactions while, there are less than 50,000 newly built properties. Based on this data, the market appears to be very much driven by resale activity. This is informative and may help with targeted marketing resource allocation.

In overall, EDA results provide important information about key predictive features, note possible defects in the data (outliers), and highlights market trends so predictive model and regression analyses can be developed on a solid framework.

Model Development and Evaluation

The dataset was divided into two portions: the training set (80%) and the testing set (20%). To equalize feature contribution to ultimately improve stability, all predictors were standardized using StandardScaler from the Python library SciKit-Learn (Pedregosa et al.,

2011). Standardizing the features was important to prevent large scale variables from dominating the model predictions.

There are three regression models developed to predict house prices: Linear Regression, Ridge Regression, and a polynomial regression with degree two. Linear Regression served as a baseline with an obvious relationship between predictors and target. Ridge Regression employed L2 regularization to the model with $\alpha = 1.0$; it was used to address potential multicollinearity feature issues. Polynomial regression would allow us to also model more complex, non-linear relationships by adding polynomial terms to the predictors. MSE and RMSE are the metrics used to estimate the efficiency and accuracy of models that characterize prediction error.

Results

Table 1: *Model Performance*

Model	**MSE (₹²)**	**RMSE (₹)**
Linear Regression	2.6561×10^{14}	16,297,635.73
Ridge Regression	2.6561×10^{14}	16,297,636.28
Polynomial (deg 2)	1.8405×10^{16}	135,665,525.31

Note: Model Performance Comparison basis MSE & RMSE

The performance comparison revealed insignificant differences between Linear regression and Ridge regression indicating that the penalty of L2 regularization at $\alpha = 1.0$ had little influence on the solution. However, the Polynomial Regression model, on the other hand, over-fitted substantially as indicated by its much higher RMSE indicating that when using advanced model families, overly complex models with little or no regularization can be risky.

In all, the evaluation suggested that linear approaches were sufficient to capture most of this data's variance indicating that house prices generally followed linear additive

relationships with predictor variables. Future performance improvements could include log-transformations to mitigate some skewness and heteroscedasticity, systematic hyperparameter optimization of Ridge Regression models, more advanced model families, such as tree-based ensembles, and also the modelling of interactions between predictor variables.

Discussion

The high performance of the linear model indicates that additive relationships do indeed account for most of the variance in the sale price in this dataset. The performance may have been improved if log-transformed the target to better manage skewness and heteroscedasticity, and then inverse transform the prediction afterwards. Hyperparameter optimization, particularly a systematic grid search to optimize Ridge's α, might improve the stability of the model even further. The usage of tree-based ensembles (for example, Random Forest and Gradient Boosting) to capture nonlinear interactions without having to deal with the combinatorial explosion of polynomial features. Residual diagnostics can be used to help endorse models assumptions, and inspire feature engineering, including, but not limited to, considering ratio features (like bath count/unit area), or grouping sparse categorical levels.

Conclusion

This lab revealed that simple regression models, when combined with thorough data cleaning and exploratory analysis, can produce unsophisticated yet relatively accurate house-price predictions. Linear Regression performed best based on the trade-off between simplicity and accuracy achieving an RMSE of ₹16.30 million. Next, focus will be on to evaluate target transformation, hyperparameter tuning, and ensemble methods to improve predictions on a skewed real-estate dataset.

References

Juhibhojani, P. (2025). *House Price Dataset*. Kaggle. https://www.kaggle.com/juhibhojani/house-price

McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 51–56. https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Waskom, M., Botvinnik, O., Hobson, P., Cole, J., Halchenko, Y., VanderPlas, J., … Quintero, E. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. https://doi.org/10.21105/joss.03021