Lab 1: Predicting House Prices Using Regression Model

Priyal Rawat

DeVos Graduate School, Northwood University

MGT 665: Solving Bus Problems W/ Machine Learning

Dr. Itauma Itauma

06/08/2025

Abstract

This paper illustrates the application of machine learning regression models to predict residential property prices, using a dataset from kaggle which includes over 180,000 property listings. Four regression models named Linear, Polynomial, Ridge, and Lasso were modelled and evaluated with using of standard metrics which are MSE, RMSE, MAE, R2. The polynomial model performed best out of the four models. Two hypotheses were tested in this study to evaluate the effect of carpet area and the number of bathrooms on the price of residential properties. The number of bathrooms was a useful predictor while carpet area was not as substantial as it was assumed Feature importance was used to explore interactions within our model. For example Furnishing Transaction had the largest overall effect. The study concluded the traditional regression methodologies provides a weak baseline for predicting housing prices, suggested the need for advanced methods for future research.

Introduction

The real estate industry has been a key focus of economic health, impacting individual wealth, local growth, and national progress from a very long period of time. Prediction of housing prices is crucial to anyone involved in or impacted by the industry, particularly for consumers like buyers or sellers, investors which are in real estate, stock market, etc., and decision-makers like government, policy makers, etc. Historically, housing price estimation has been reliant on appraisals and some historical trends data, which can be problematic due to their lack in subjectivity and scalability (Zhang et al., 2021). The recent increases in structured data availability and means of analysis with machine learning have made predictive modeling an increasingly viable option for stakeholders and actors in the real estate industry (Sreelekshmi et al., 2020).

Machine learning models are well suited and have the ability, as they can uncover or reveal patterns or interdependencies between many housing variables and house price (Anisotropic, 2018). The study will aim to utilize and compare different regression techniques to predict house prices based on various structural and locational features. It focuses on regression models which are Linear Regression, Polynomial Regression, Ridge Regression, and Lasso Regression, with the goal of determining the best model to create generalizable and accurate predictions. The study will factor in various features in a dataset including carpet area, number of bathrooms, floor, property status, number of parking spots, and other features.

This paper is focused on building a predictive framework that is supported by data preprocessing, exploratory data analysis (EDA), and comprehensive model evaluation using an appropriate metric performance indication. The final product is a comparative analysis indicating that while the research asks which model performs better for this data, it also aims to indicate which features are most important and what areas require additional research or optimization.

Literature Review

For a long time now, the prediction of housing prices has been an important field of study, due to its economic impact and the consequences for buyers, sellers, developers, and policy makers. Many studies have conducted analysis in the area and utilized statistical and machine learning methods, stressing feature selection and suitable algorithms. The existing literature provides a good starting point, this study seeks to go further by applying machine learning techniques to a large, diverse housing dataset representative of the Indian real estate market, a field that has increasingly been gaining attention in current research. The present study hopes to fill this gap, through the implementation of regression-based machine learning models to a rich dataset from Kaggle, that includes a range of diverse features meaningfully representing the Indian housing market.

Zhang et al. (2021) noted that Multiple Linear Regression (MLR) is a good benchmark method based on it is often straightforward to create and allows for easy explanation. They showed that super area, carpet area, number of bathrooms and property status are worth noting in regard to explaining price. At the same time, they recognize that linear models struggle with multicollinearity and nonlinear relationships, which can lead to poor results against complex datasets. Therefore, in line with Zhang et al. (2021), it was important to create models that also contained Polynomial Regression. Sreelekshmi et al. (2020) demonstrated that polynomial models are highly useful as they can describe nonlinear trends in housing prices and datasets that do not imply linear relationships.

Feature selection is also a key component of successful predictive modeling. Tripathi and Jain (2021) indicate that traditional datasets in housing price research have input variables that were limited and also omitted contextual categorical variables, relying on attributes such as furnishing status, type of transaction, facing and ownership type, all variables that help provide vital context within certain markets. The study includes numerical feature variables, as well as categorical features, with a preprocessing pipeline that encodes, imputed and normalizes variables. The additional data provided by categorical properties decreases the predictive depth of the input data.

Regularized regression models were created to address the issues of multicollinearity and overfitting. Specifically, Ridge regression and Lasso regression (Tripathi & Jain, 2021;

Anisotropic, 2018) were selected due to their demonstrated ability to improve model generalizability with high dimensional datasets. Ridge regression was an ideal choice when trying to deal with correlated predictors since it apportions weight among the coefficients instead of eliminating predictors completely. Alternatively, Lasso regression accounts for multicollinearity and performs feature selection, by reducing the insignificant feature coefficients to zero (Zhang et al. 2021). The complementary attributes of Ridge and Lasso regression justify their inclusion from a comparative model perspective.

In addition to earlier research, this paper takes inspiration from existing research and focuses on applying machine learning regression models to analyze a unique and different real-life large scale Indian housing dataset containing more than 180,000 records. This study also expanded the assessment by utilizing four strong performance evaluation metrics which are MSE, RMSE, MAE, and R^2, are all of equal weight; this is supported by prior studies (Agarwal et al., 2023; Sreelekshmi et al., 2020). The models were compared collectively based on the models' ability to capture large error, error in the wrong direction, and error that is explained quantitatively by the model.

To sum up, this research builds on existing techniques while addressing a clear research gap in the economic literature, conducting a similar analysis on a large Indian housing dataset using a broader set of categorical and numerical features. It assesses traditional and regularized models while providing evidence of their strengths and weaknesses in a dynamic and developing real estate market.

Methodology

Dataset Description

The dataset utilized in this analysis was taken from Kaggle and consisted of an entire history of residential property listings including 187,531 rows and 21 columns of data. The variables columns used are carpet area, property status, floor, transaction type, furnishing, facing, overview, name of society, number of bathrooms, number of balconies, car parking, ownership type, super area, dimensions, plot area, property title, total amount, price per square foot, and location.

The feature like Price is in INR which represents the target variable for the predictive modeling project. There are a few columns such as Carpet Area and Plot Area with missing values, and others such as Amount in INR with non-numeric value. The dataset reflects listings from the larger cities in India, including New Delhi, Bangalore and Mumbai, as well as regional cities.

Hypothesis

The analysis was guided by the formulation of two hypotheses before beginning model development. The first hypothesis (H1) stated that larger carpet areas significantly increase house prices. This hypothesis was based upon the presumption that buyers placed a premium on usable living space. This was supported by Zhang et al. (2021) who found that carpet area was identified as a structural characteristic that was one of the major determinants of price in housing markets. The second hypothesis (H2) stated that the number of bathrooms was a very strong predictor of house price. The assumption made with this hypothesis is that bathrooms improves comfort, utility, and increases with property size and quality. Anisotropic (2018) and Sreelekshmi et al. (2020) stated that a count of bathrooms were indeed a very important feature with significant influence in most machine learning models when predicting home prices. The hypotheses were measurable by evaluating the feature importance.

Data Preprocessing

This study utilized an approach based on supervised machine learning that involved developing predictive models for estimating prices of residential properties. The raw dataset that was used in this study was named house_prices.csv which was uploaded into the Google Colab and reviewed for completeness. A missing value analysis was performed on the dataset to identify columns that had substantial amounts of missing values. The two columns with 100% missing values, "Plot Area" and "Dimensions," were dropped. In addition, any rows that presented missing values for the target variable Price were also dropped.

For the rest of the columns, for numerical missing values, based on variability in the data, the median was used for imputation. For categorical missing values, the mode value was used. This method was chosen so that data integrity was kept in place, while avoiding outlier bias a fuller dataset. All categorical variables were encoded via Label Encoding to ensure they were all in the appropriate numerical forms that a machine could read.

Exploratory Data Analysis (EDA)

An exploratory data analysis (EDA) was performed in order to better understand the layout of the data set, understand critical features that affect housing prices, identify outliers or anomalies. Histograms were generated to plot distribution of target variable Price in rupees represented with a right-skewed distribution which is typical of real estate data. The correlation coefficient heatmap visually expresses correlations, including negative and positive, among numeric variables. The heatmap demonstrates strong predictors such as Super Area and Carpet Area.

Boxplots were used to examine how categorical variables such as Transaction Type and Furniture affect property prices. The boxplots helped communicate the median price for each category of the categorical features. Additional pairplots were generated to illustrate pairwise interaction between features, and a joint grid histogram provided visibility of all numerical feature distributions. These visualizations were helpful in creating a plan of feature selection and appropriate modeling.

Model Development

Four regression techniques were utilized to predict housing prices: Linear Regression, Polynomial Regression, Ridge Regression, and Lasso Regression. In considering the models to create, the selection was based on literature to identify the advantages of each. Linear Regression serves as the baseline regression model, due to its simplicity and interpretability (Zhang et al, 2021). Polynomial regression is useful for obtaining nonlinear relationships that may not be captured with a linear model (Sreelekshmi et al, 2020). Ridge and Lasso regression presented regularization techniques that avoid overfitting and excellent for multicollinearity thas may typically occur in real estate datasets (Tripathi & Jain, 2021).

All models were trained with an 80:20 train-test split of the dataset. The feature set (X) does not include the target variable (Price), and the models were executed using Python libraries such as scikit-learn and numpy.

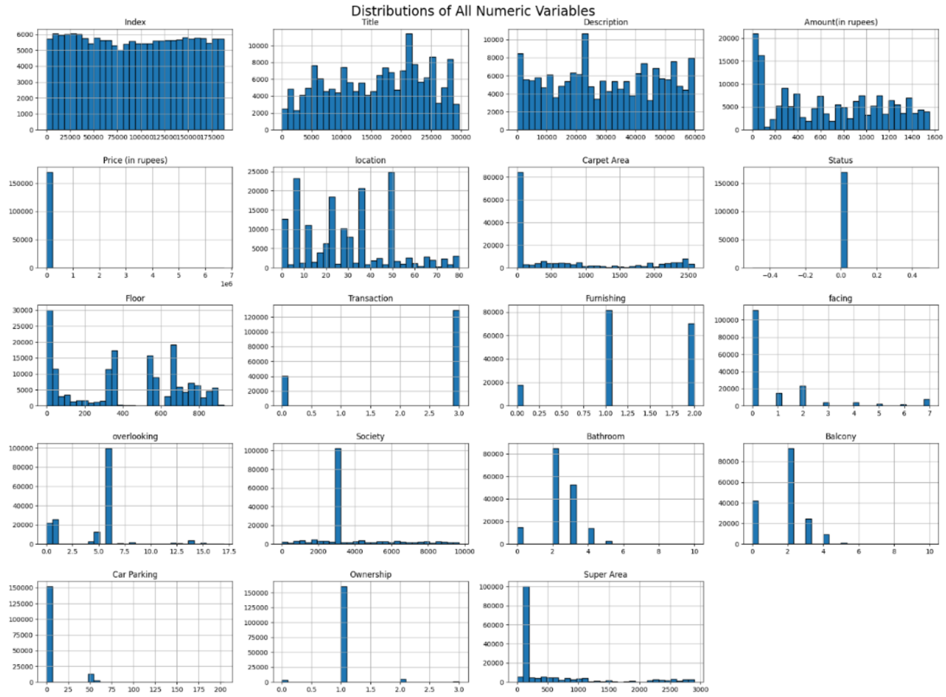
Model Evaluation

For this study, the model performance was assessed through four regression metrics which are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R2 Score. MSE is an average of the squared differences between the actual and predicted values, which tends to penalize larger errors our predictions made even more. RMSE is the square root of the value of MSE, which allows us to measure the MSE in the same units as the values in the target variable, leading to an improvement in interpretability and especially valuable in the context of real estate. MAE is the average of absolute errors in predictions, which does not consider direction, and it is less sensitive to outliers. R2 Score assesses the proportion of variance in the dependent variable explained by the model, which is a measure of goodness-of-fit and generalizability. These four metrics together provide a balanced consideration of how precise, robust and explanatory the model is. The results of each model were plotted into a summary table to compare and find the best-performing algorithm.

Results

Exploratory Data Analysis (EDA)

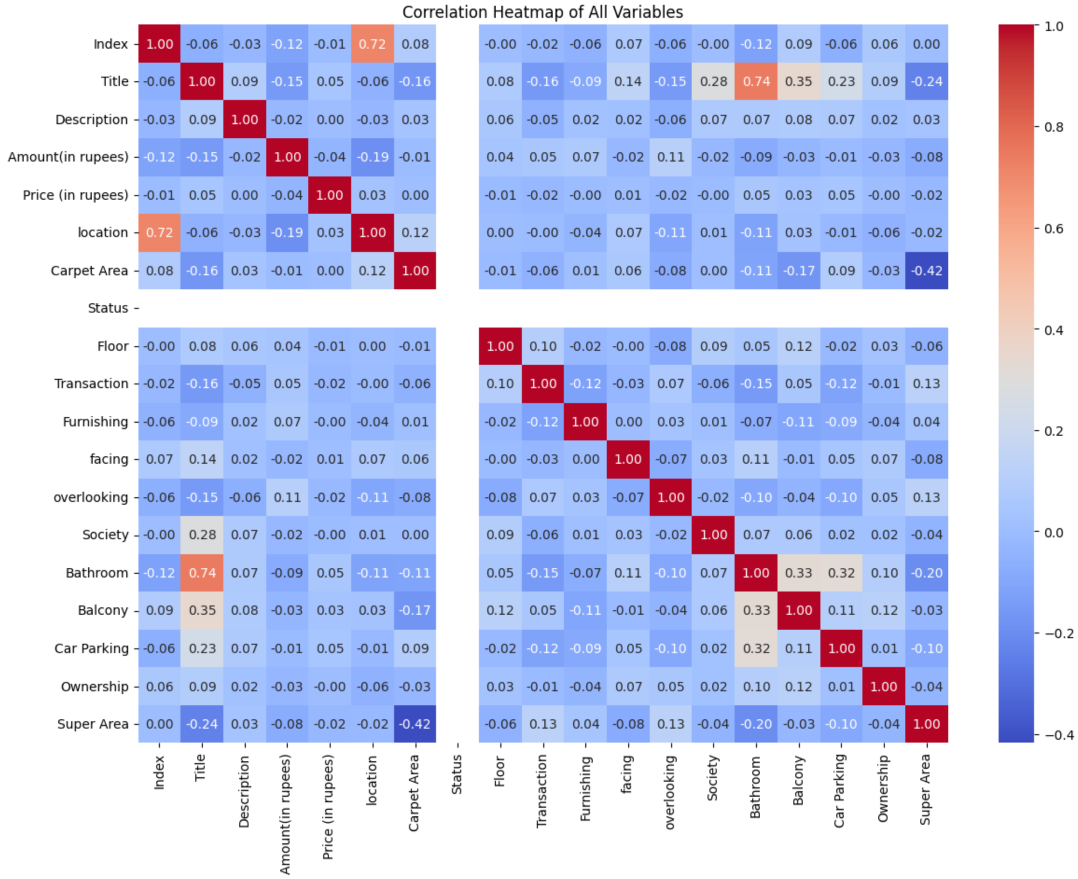
Figure 1: *Distribution of Numeric Variables*



Note: The combined histogram plot provides a summary of the distribution patterns across all numeric features in the dataset. This histogram shows the distribution of house prices in the dataset. Most properties are priced at the lower end of the spectrum. The sum of the values for Price, Carpet Area, Super Area and other variables all show a strong right skewness and therefore extreme high values or outliers.

The amounts of the Bathroom, Balcony and Car Parking columns show a high concentration of counts for lower amounts, suggesting that most properties had a low count for these features. The numeric columns that behave like categorical columns such as Furnishing, Transaction, and Ownership, show highly unbalanced and varied distributions and can be thought of as limited variation across the dataset and are of further interest to investigate. The summary provides evidence that the dataset is highly heterogeneous, and the skewed and sparse features strongly suggest the need for data normalization and advanced modeling algorithms to ensure prediction accuracy.

Figure 2: *Correlation Heatmap of All Variables*

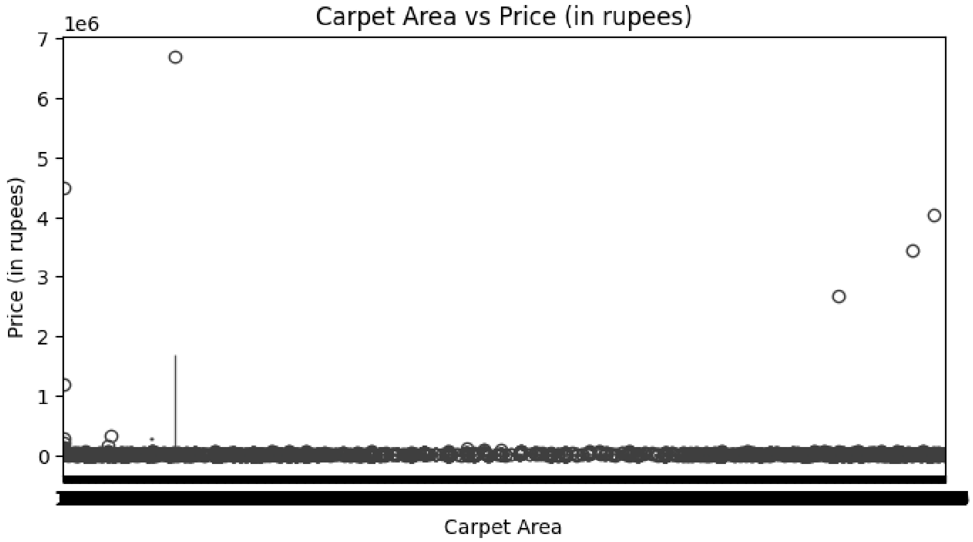


	Index	Title	Description	Amount(in rupees)	Price (in rupees)	location	Carpet Area	Status	Floor	Transaction	Furnishing	facing	overlooking	Society	Bathroom	Balcony	Car Parking	Ownership	Super Area
Index	1.00	-0.06	-0.03	-0.12	-0.01	0.72	0.08		-0.00	-0.02	-0.06	0.07	-0.06	-0.00	-0.12	0.09	-0.06	0.06	0.00
Title	-0.06	1.00	0.09	-0.15	0.05	-0.06	-0.16		0.08	-0.16	-0.09	0.14	-0.15	0.28	0.74	0.35	0.23	0.09	-0.24
Description	-0.03	0.09	1.00	-0.02	0.00	-0.03	0.03		0.06	-0.05	0.02	0.02	-0.06	0.07	0.07	0.08	0.07	0.02	0.03
Amount(in rupees)	-0.12	-0.15	-0.02	1.00	-0.04	-0.19	-0.01		0.04	0.05	0.07	-0.02	0.11	-0.02	-0.09	-0.03	-0.01	-0.03	-0.08
Price (in rupees)	-0.01	0.05	0.00	-0.04	1.00	0.03	0.00		-0.01	-0.02	-0.00	0.01	-0.02	-0.00	0.05	0.03	0.05	-0.00	-0.02
location	0.72	-0.06	-0.03	-0.19	0.03	1.00	0.12		0.00	-0.00	-0.04	0.07	-0.11	0.01	-0.11	0.03	-0.01	-0.06	-0.02
Carpet Area	0.08	-0.16	0.03	-0.01	0.00	0.12	1.00		-0.01	-0.06	0.01	0.06	-0.08	0.00	-0.11	-0.17	0.09	-0.03	-0.42
Status																			
Floor	-0.00	0.08	0.06	0.04	-0.01	0.00	-0.01		1.00	0.10	-0.02	-0.00	-0.08	0.09	0.05	0.12	-0.02	0.03	-0.06
Transaction	-0.02	-0.16	-0.05	0.05	-0.02	-0.00	-0.06		0.10	1.00	-0.12	-0.03	0.07	-0.06	-0.15	0.05	-0.12	-0.01	0.13
Furnishing	-0.06	-0.09	0.02	0.07	-0.00	-0.04	0.01		-0.02	-0.12	1.00	0.00	0.03	0.01	-0.07	-0.11	-0.09	-0.04	0.04
facing	0.07	0.14	0.02	-0.02	0.01	0.07	0.06		-0.00	-0.03	0.00	1.00	-0.07	0.03	0.11	-0.01	0.05	0.07	-0.08
overlooking	-0.06	-0.15	-0.06	0.11	-0.02	-0.11	-0.08		-0.08	0.07	0.03	-0.07	1.00	-0.02	-0.10	-0.04	-0.10	0.05	0.13
Society	-0.00	0.28	0.07	-0.02	-0.00	0.01	0.00		0.09	-0.06	0.01	0.03	-0.02	1.00	0.07	0.06	0.02	0.02	-0.04
Bathroom	-0.12	0.74	0.07	-0.09	0.05	-0.11	-0.11		0.05	-0.15	-0.07	0.11	-0.10	0.07	1.00	0.33	0.32	0.10	-0.20
Balcony	0.09	0.35	0.08	-0.03	0.03	0.03	-0.17		0.12	0.05	-0.11	-0.01	-0.04	0.06	0.33	1.00	0.11	0.12	-0.03
Car Parking	-0.06	0.23	0.07	-0.01	0.05	-0.01	0.09		-0.02	-0.12	-0.09	0.05	-0.10	0.02	0.32	0.11	1.00	0.01	-0.10
Ownership	0.06	0.09	0.02	-0.03	-0.00	-0.06	-0.03		0.03	-0.01	-0.04	0.07	0.05	0.02	0.10	0.12	0.01	1.00	-0.04
Super Area	0.00	-0.24	0.03	-0.08	-0.02	-0.02	-0.42		-0.06	0.13	0.04	-0.08	0.13	-0.04	-0.20	-0.03	-0.10	-0.04	1.00

Note: The correlation heatmap shows correlations between each numerical variable in the dataset. Nearly all the variables show weak to very weak correlation with the price variable, which target variable is labeled Price in rupees. Notably, variables such as Carpet Area, Bathroom, Super Area all have slight negative or near-zero correlations which suggest a limited linear relationship.

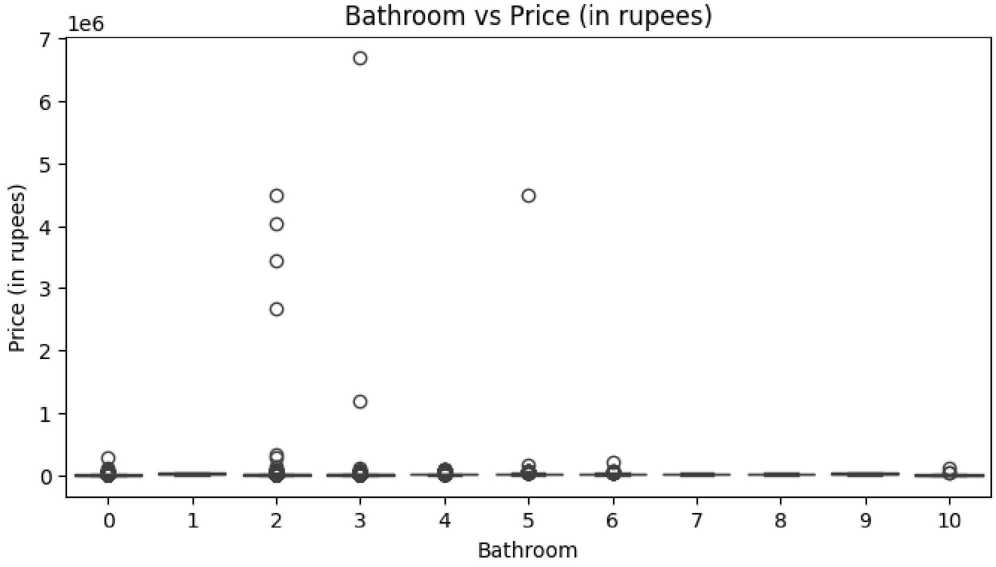
There is a moderate correlation between the variables Bathroom and Title, and between Bathroom and Balcony, indicating that these are features that often covary. In summary, all the correlation values are low suggesting that price prediction in this dataset probably cannot be characterized by simple linear trends, which will reinforce the necessity to use a more flexible approach/models like Polynomial Regression, Ridge Regression or Lasso Regression.

Figure 3: *Boxplot of Carpet Area vs Prices*



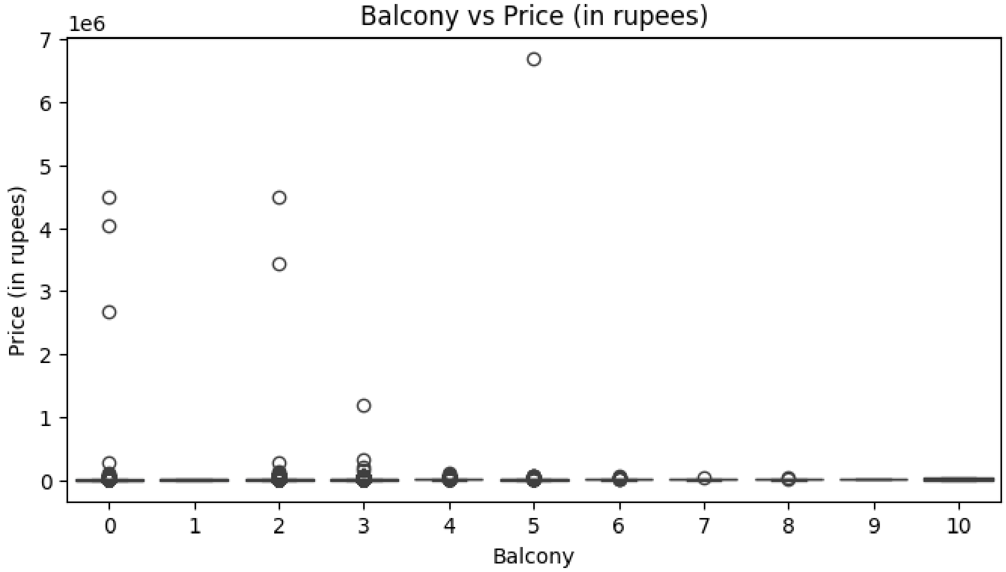
Note: The box plot shows slight increasing trend, larger carpet area and increased price. However, there is a large cluster of low prices and areas and on the other end there are a few outliers with very large carpet areas and high price which supports positive but non-linear correlation.

Figure 4: *Boxplot of Bathroom Vs Price*



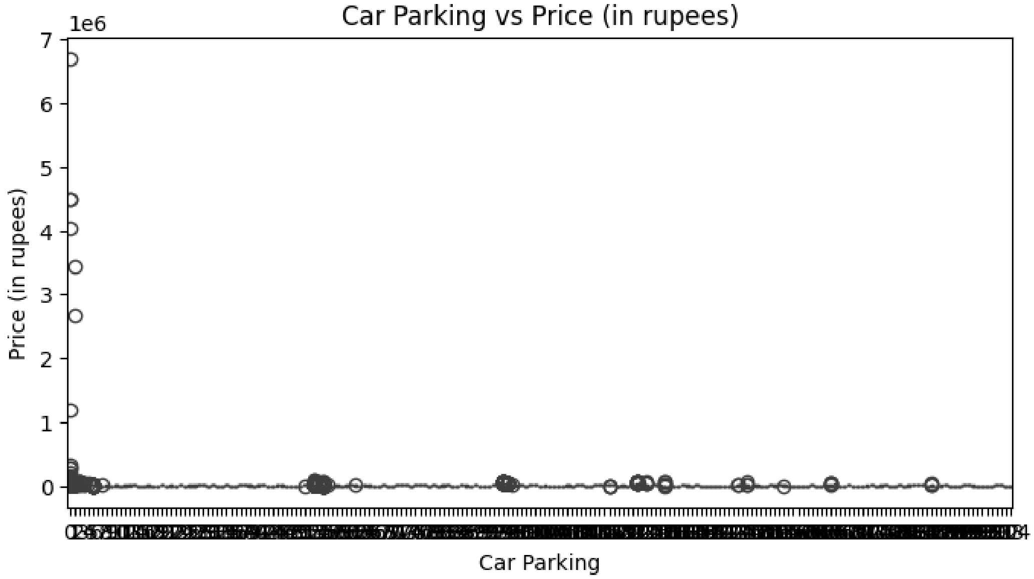
Note: The plot shows most of the properties cluster around the low end of the range of 1-3 bathrooms and there was a general trend that associated higher priced listings with more bathrooms, however they do not all have increased prices. The high price listings are scattered throughout the bathroom count and do not define particular pattern.

Figure 5: *Boxplot of Balcony Vs Price*



Note: As noted with bathrooms this plot shows limited differentiation in price based on the number of balconies. There are many listings with 0-3 balconies, and there was not much of a promising trend that related price to balcony count suggesting weak correlation.

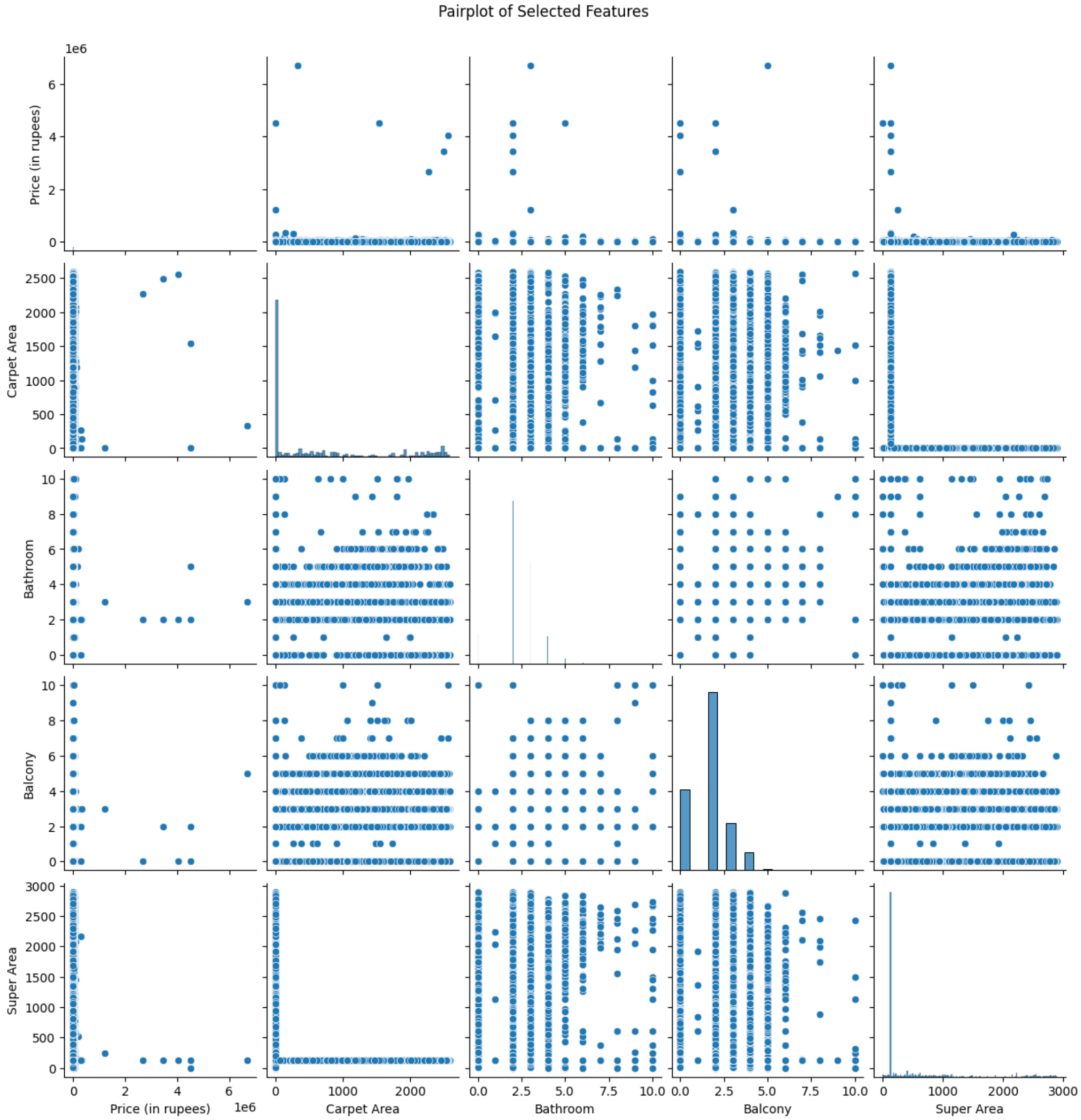
Figure 6: *Boxplot of Car Parking vs Price*



Note: The uncertainty in relationship between car parking availability and house price. The chart shows that for each of the car parking categories there is a broad range of prices with large clusters on the low end of the price, and few higher priced outliers suggesting car parking may not be a strong predictor of the price when analyzed in isolation.

Overall, these visualizations suggest that while features such as carpet area and bathrooms show a mild positive relation to pricing, balconies and parking have weaker relations, indicating the need for models that can capture more complex interactions between multiple features and non-liner effects.

Figure 7: *Pairplot of Selected Features*



Note: The pairplot of selected features provides a multidimensional overview of relationships between key numerical variables which are Price, Carpet Area, Bathroom, Balcony, and Super Area.

Each diagonal plot depicts the distribution of individual features, whereas the scatter plots highlight pairwise interactions. The plot indicates the majority of data points are clustered around lower values with a few extreme outliers existing at the higher ends of price, carpet area, and super area. There is a small positive trend along the Carpet Area and Super Area scatter plot indicating that larger homes generally have larger usable spaces. The scatterplot involving Price vs Carpet Area and the Price vs Super Area also show sparse but observable upward trends indicating that units with larger areas tend to have higher prices. Bathroom and Balcony features appear discretely spaced with slight differences in price suggesting these features had limited impact on cost. The lack of strong widespread linear relationships for most variable pairs also facilitates the use of complex regressions that can better account for more complicated non-linear relationships between housing features.

Model Performance

Table 1: *Performance Metrics for Regression Models*

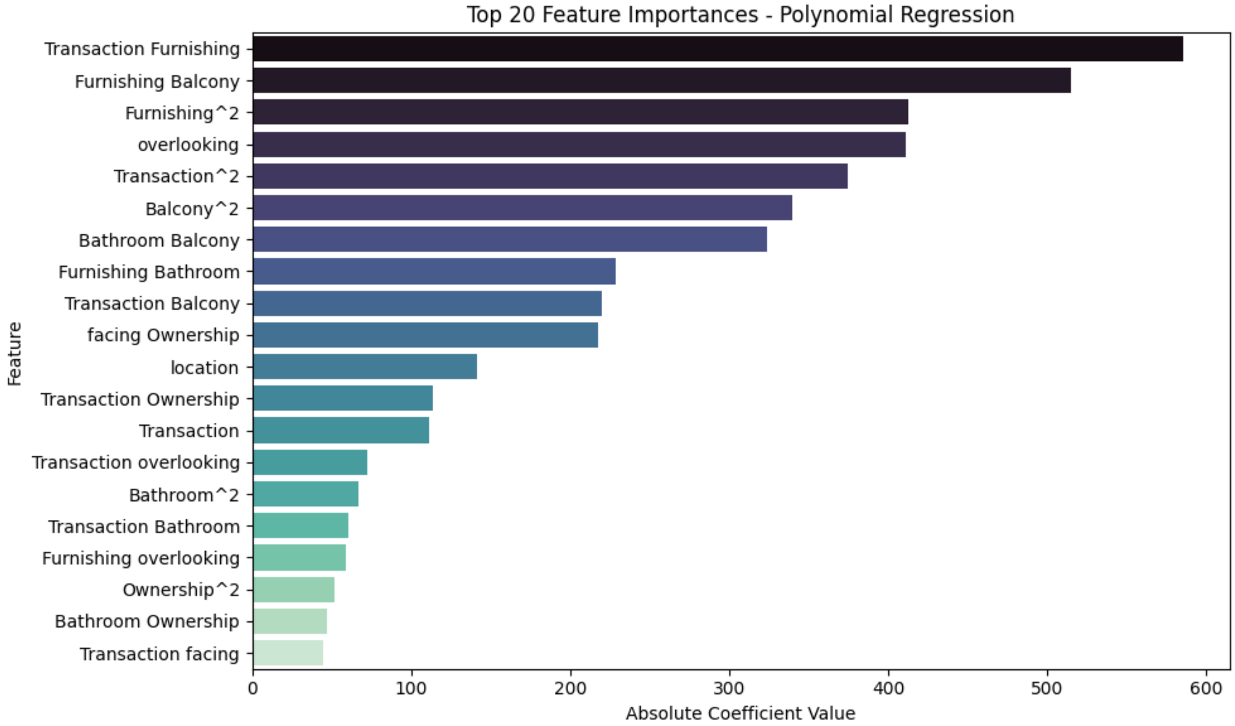
Model Name	MSE	RMSE	MAE	R2 Score
Polynomial Regression	1970632324.11	44391.8	2777.54	0.0041
Linear Regression	1972067003.71	44407.96	3183.4	0.0034
Ridge Regression	1972067009.22	44407.96	3183.4	0.0034
Lasso Regression	1972067347.99	44407.96	3183.49	0.0034

Note: Table 1 shows the performance measures of the four regression models, Polynomial Regression, Linear Regression, Ridge Regression, and Lasso Regression, that were utilized on the task of predicting house prices. Performance on the models were measured according to Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and the R2 score.

Among these four regression models, Polynomial Regression had the best performance, with the lowest MSE of 1,970,632,324.11, the lowest RMSE of 44,391.8 and the lowest MAE of 2777.54, and highest R2 score of 0.0041. Although all R2 values for all models were low and clearly not very explanatory, the consistently lower error values for Polynomial Regression would make it the best model overall, in terms of minimizing prediction error. This is further substantiated by the model being in a position to account for non-linear relationships, which are likely to be present in a real-world dataset like this with complexity and which is large scale. Conclusively, Polynomial Regression was the best performing model of the above.

Feature Importance

Figure 8: *Feature Importance*



Note: The bar chart illustrates the 20 most impactful features derived from the Polynomial Regression model's assessment of house prices. Features in the model are original and polynomial features i.e., squared terms and interaction of features. Each bar captures the absolute value of the coefficient for a feature, which corresponds to its direct influence on the target variable - house price.

The most impactful features include interaction terms such as Transaction Furnishing, Furnishing Balcony, Furnishing squared. These features include both interaction terms and squared terms which capture evidence that shows the relationships between property features and price are not linear and depend on how features interacted with each other to influence housing price. For example, a furnished resale property is likely to generate different pricing than an unfurnished version of the same property. Balcony squared, Bathroom Balcony, and Transaction squared are also strong examples of nonlinear influence and indicate that the impact of small increases in the number of features can generate incremental price influences that multiply to great effects.

Discussion

The results of this analysis suggest that that despite four regression models Linear, Polynomial, Ridge, and Lasso producing almost identical error metrics, their R2 values, or percentage of the variance explained in house prices, were very low (as shown in Table 1). Each regression model's R2 were all less than 0.005 and therefore indicated that none of the model four models were able to adequately represent and explain the complex nature of the relationships embedded in the real estate pricing trends. These negative outcomes were unproductive and negated the initial characteristics of our housing dataset where it was expected feasible characteristics followed by positive features and strong predictive capabilities using traditional regression modelling approaches.

Among the models analyzed, Polynomial Regression demonstrated slightly higher statistics in Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) which indicates not only that the relationships are nonlinear with interaction effects, but there is a degree of complexity that simpler models cannot account for, with small overall improvements. In other words, it appears that the models had a more difficult time estimating luxury or outlier properties, not only due to nonlinear relationships but possibly because of omitted outside variables such as prestige of locality, quality of view or builder brand.

To assess and validate assumptions, two hypotheses were formed prior to building a model, our H1 was that larger carpet area significantly increases house price, and H2 that number of bathrooms is a strong predictor of house price as it is representative of comfort and house size.

To test these hypothesis, analysis was done using the Polynomial Regression feature importance with squared terms and feature interaction terms. Notably, Bathroom ranked quite high, due to interactions such as Bathroom Balcony and Bathroom2, but Carpet Area did not appear in the top 20 strong predictors. This may indicate that just usable floor space will not drive price alone, but how the usable floor interacts with other features such as type of furnishing, type of transaction or owner status. It was surprising for us that the majority of normal top predictors were interaction features of features such as Transaction Furnishing, Furnishing Balcony and Furnishing2 suggesting that price determinants is complicated and non-linear.

These findings challenge traditional assumptions and highlight that price is not purely based on only simple attributes, and the way features relate or interact with each other matters. For example, a furnished resale property may sell for more than an unfurnished property of the same quality overall condition, but only when certain combinations like balcony, bathroom etc. are met or fulfilled.

In summary, this analysis has showed that linear and regularized regression approaches may provide useful baselines for model comparison, but it is difficult to analyze real estate price behavior in rich, diverse, datasets. It also emphasized that certain assumptions must be validated factually since even intuitive domain knowledge does not always align with factual statistical data.

Conclusion

The objective of this research was to develop an understanding of housing prices using a comprehensive dataset of housing attributes, and utilizing 4 regression models, Linear, Polynomial, Ridge, and Lasso. All four models produced somewhat reasonable or acceptable predictions of housing prices, however, with modeling in housing it once again highlighted the importance of considering nonlinearities and interaction terms, since the polynomial performed once again just a notch above the others.

The original hypothesis of the research assumed that carpet area and bathrooms would be considered important predictors of price. In the end bathrooms surfaced as the top feature based on its interactions, but carpet area did not demonstrate influence in the way it was assumed. Rather, the interaction terms, Furnishing Transaction Type and Furnishing Balcony, were significantly more important than anticipated based on the polynomial feature importance analysis. This demonstrates that housing prices are influenced based not only on single attributes of data, but how all the attributes together may interact and combine.

Even with stringent pre-processing, modeling and evaluating processes in place the low R2 values indicate that traditional regression techniques are not adequate in what appears to be a complex model of housing price behaviors, especially given the diversity and extent of the data. There are likely additional factors besides value added that could account for price variance, such as aspects related to location quality, builder reputation, included amenities, and dynamics with foreign markets, all of which were not available in this dataset.

In conclusion, this study demonstrated the importance of hypothesis testing, non-linear modeling capabilities, and that using data for validation should be prioritized over intuition alone. A next step in this area would be to increase the number of location-based, economic, and amenity-based variables while using more advanced ensemble modelling based techniques e.g., Random Forest, XGBoost, CatBoost. These modelling techniques will allow for more nonlinear modeling capabilities and feature interaction modeling without the need to hand craft predictive feature transformations.

References

Zhang, Y., Liu, H., Zhang, Q., & Tian, Y. (2021). Housing price prediction based on multiple linear regression. *Scientific Programming, 2021*, 1–8. https://doi.org/10.1155/2021/5520032

Sreelekshmi, T. R., Vishnupriya, M. S., & Anjusha, V. S. (2020). House price prediction using machine learning techniques. *Procedia Computer Science, 172*, 300–305. https://doi.org/10.1016/j.procs.2020.05.043

Tripathi, A., & Jain, R. (2021). Predictive modeling in the real estate sector using regularized regression. *Procedia Computer Science, 185*, 315–322.

Agarwal, A., Gupta, A., & Nayak, D. (2023). Data science-driven real estate analytics. *Asian Journal of Research in Computer Science, 13*(3), 56–66.

Anisotropic. (2018). Predict house prices with machine learning. *Towards Data Science*. https://towardsdatascience.com/predict-house-prices-with-machine-learning-5b475db4e1e/