Laboratory Exercise # 1: India House Rent Price Prediction Model Using Linear Regression

Joshua Kyle K. Entrata
CSELEC2C - Machine Learning

I. INTRODUCTION

Housing in India is diverse as is its culture, varying from palaces of erstwhile maharajas to modern apartment buildings in big cities to tiny huts in far-flung villages. The houses are around 18% of the world's population in 2.4% of the world's land area.[1] As the world's most populated country, it might face a perpetual housing crisis. The demand for better living spaces has been booming for the past few years in their housing sector.

Predicting rental prices holds significant value both for the real estate market and individuals. It can be used as a guide for pricing strategies to maximize the profits while ensuring competitiveness. This balance between supply and demand benefits everyone involved because of a more transparent and efficient market.

As a Data Scientist working for a real-estate company that is providing leases and letting for customers, my objective is to develop a model that accurately predicts the renting price of properties based on information from housing data.

To achieve this objective, linear regression was used. It is a statistical method that predicts the value of a variable (dependent variable) based on the value of another value (independent variable) by fitting a linear equation to observed data. The simplicity and interpretability of this data analysis technique make it a powerful tool for predicting outcomes based on linear relationships between variables.

[1] Marriner, Katie. "India is overtaking China today as the world's most populous country – according to this projection". *MarketWatch*.

Linear regression will serve as the backbone of the predictive model by analyzing the relationship between various property characteristics, such as size, floor location, and BHK, and their renting prices. This approach helps in giving insights into the factors that impact the renting prices in India's housing market.

II. METHODOLOGY

2.1 Data Loading and Exploration

The dataset used for the prediction model consists of 4,746 rows, detailing 12 unique property characteristics, including the date the listing was posted, total number of bedroom, hall, and kitchen (BHK), rent cost, property size, floor location, area type, area locality, city, furnishing status, tenant preferred, bathroom, and point of contact.

	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
0	2022-05-18	2	10000	1100	Ground out of 2	Super Area	Bandel	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner
1	2022-05-13	2	20000	800	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
2	2022-05-16	2	17000	1000	1 out of 3	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
3	2022-07-04	2	10000	800	1 out of 2	Super Area	Dumdum Park	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner
4	2022-05-09	2	7500	850	1 out of 2	Carpet Area	South Dum Dum	Kolkata	Unfurnished	Bachelors	1	Contact Owner
...	...	...	...	...	...	...	...	...	...	...	...	...
4741	2022-05-18	2	15000	1000	3 out of 5	Carpet Area	Bandam Kommu	Hyderabad	Semi-Furnished	Bachelors/Family	2	Contact Owner
4742	2022-05-15	3	29000	2000	1 out of 4	Super Area	Manikonda, Hyderabad	Hyderabad	Semi-Furnished	Bachelors/Family	3	Contact Owner
4743	2022-07-10	3	35000	1750	3 out of 5	Carpet Area	Himayath Nagar, NH 7	Hyderabad	Semi-Furnished	Bachelors/Family	3	Contact Agent
4744	2022-07-06	3	45000	1500	23 out of 34	Carpet Area	Gachibowli	Hyderabad	Semi-Furnished	Family	2	Contact Agent
4745	2022-05-04	2	15000	1000	4 out of 5	Carpet Area	Suchitra Circle	Hyderabad	Unfurnished	Bachelors	2	Contact Owner

4746 rows × 12 columns

Figure 1. *Initial House Rent Dataset*

Initial data exploration was crucial for understanding the structure and can help in deciding strategies to create the model. Checking the data types and unique values in categorical columns helped in understanding the impact of these characteristics to the pricing of the houses.

2.2 Preprocessing

2.2.1 Feature Engineering. The 'Posted On' date was transformed into a new feature, 'Age of Listing (days)', which helped the machine learning model to learn from the data more effectively. Additionally, the 'Floor' information was split into two separate features, 'Floor Number' and 'Total Floors', to get a better understanding of the impact of the house's elevation on its rent price. This phase resulted in a refined feature set, although 'Point of Contact', 'Area Locality', and 'Tenant Preferred' were removed for their lower relevance to the renting price.

2.2.2 Feature Selection. The objective was to construct a predictive model that could accurately determine the rent prices based on variables that may have a direct or indirect impact. But the priority is to include attributes with a direct impact on rental prices. Through careful analysis, selected variables were included in the feature set to help in the accuracy of the prediction model. These features consist of 'BHK', 'Size', 'Floor Number', 'Total Floors', 'Area Type', 'City', 'Furnishing Status', and 'Bathroom'. Features like 'BHK' and 'Bathroom' were included because of their strong correlation with price variations, while others, such as 'Area Type' and 'City', were included because it shows the desirability aspect of the houses. This selection allowed the model to focus on variables that significantly affect rental pricing to increase its accuracy.

	BHK	Rent	Size	Area Type	City	Furnishing Status	Bathroom	Age of Listing (days)	Floor Number	Total Floors
0	2	10000	1100	Super Area	Kolkata	Unfurnished	2	54	0	2
1	2	20000	800	Super Area	Kolkata	Semi-Furnished	1	59	1	3
2	2	17000	1000	Super Area	Kolkata	Semi-Furnished	1	56	1	3
3	2	10000	800	Super Area	Kolkata	Unfurnished	1	7	1	2
4	2	7500	850	Carpet Area	Kolkata	Unfurnished	1	63	1	2
...	...	...	...	...	...	...	...	...	...	...
4741	2	15000	1000	Carpet Area	Hyderabad	Semi-Furnished	2	54	3	5
4742	3	29000	2000	Super Area	Hyderabad	Semi-Furnished	3	57	1	4
4743	3	35000	1750	Carpet Area	Hyderabad	Semi-Furnished	3	1	3	5
4744	3	45000	1500	Carpet Area	Hyderabad	Semi-Furnished	2	5	23	34
4745	2	15000	1000	Carpet Area	Hyderabad	Unfurnished	2	68	4	5

4746 rows × 10 columns

Figure 2. *An overview of the features included in selection*

2.2.3 One-Hot Encoding. Numeric features such as 'BHK', 'Size', 'Bathroom', 'Age of Listing (days)', 'Floor Number', and 'Total Floors' required no further manipulation since they are already in a numeric format. In contrast, categorical variables underwent one-hot encoding, transforming each string value into its own separate boolean column to help in the model's learning process.

2.2.4 Splitting the Dataset. The dataset was divided into two parts: a training set and a test set. With this split, 20% of the data was allocated for testing and performed with a consistent random state to ensure reproducibility of the results across multiple runs. Given the size and characteristics of the dataset, a test size of 0.2 is reasonable and is commonly used in machine learning.

2.2.5 Feature Transformation. To address the complexity of how property attributes like size and location influence renting prices, a preprocessing method called PolynomialFeatures was employed to help understand the complex relationships among the features. For example, a large house in the main city might have a higher renting price than a larger one far from the city. Using this method can help the model in understanding how everything is connected, which can lead to a better accuracy of rent prices based on its characteristics.

2.3 Model Training

After the preparation of the data, the next step is to train the predictive model using linear regression, a powerful statistical method for modeling relationships between variables. The training data was fitted to this model to develop a formula that accurately predicts renting prices from the transformed polynomial features of properties.

2.4 Model Evaluation

Finally, the performance of the model using the testing set was evaluated. This evaluation helps in understanding how well the method can predict renting prices for new, unseen properties. There are two key metrics that were used for this purpose: the Mean Squared Error (MSE) and the Coefficient of Determination (R^2 score). The MSE measures the average squared difference between the actual and predicted rent prices, which provides an insight into the accuracy of the predictions. The R^2 score measures how well the regression predictions approximate the real data points. An R^2 of 1 indicates a perfect prediction.

III. EXPERIMENTS

3.1 Feature Engineering

3.1.1 Transformation of 'Posted On' Date. Analyzing the duration of a property that it has been listed for sale or rent can offer valuable insights into predicting the market value of the properties. This transformation from a simple listing date to number of days can help in analytical purposes. For example, longer-listed properties might indicate a low demand, which could impact its value to force the landlord to adjust the rent by lowering it to attract potential tenants or buyers.

To compute the 'Age of Listing', the latest date in the dataset was selected as the reference for the present date. This was used to calculate the age of listing by subtracting the 'Posted On' date of each property from this latest date in the dataset.

By transforming date values to a numeric format, it allows for more straightforward incorporation of variables into regression models.

A histogram was used to understand the distribution of 'Age of Listing (days)'. This can help in visualizing the most common age range of listings.

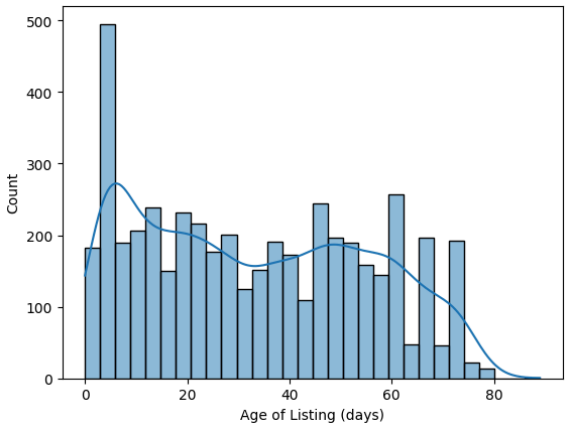

Figure 3. *Distribution of Age of Listing (days)*

3.1.2 Splitting 'Floor' Information. The floor on which a property is located can significantly influence its value and its price. In the dataset, the floor information for houses was listed as a single attribute. To increase the accuracy of the predictive model, the data was split into two distinct features: 'Floor Number' and 'Total Floors'.

An initial exploration of the data was done to see if there are non-integer values beside the " out of " string. For the first part of the split, three string values were detected: 'Ground', 'Upper Basement', and 'Lower Basement'. The floor numbering in the majority of European countries considers that the ground floor has no number or 0, the next floor up is assigned the number 1 and is the first floor, and the first basement level gets -1, and so on[2]. That is why a function was created to assign the floor to a corresponding integer. In case of the property having no other floors, the value of the floor number was just assigned to the total floors.

During the split, there were instances wherein the floor number values were greater than total floors. That is why it is assumed that it was an input error and was just switched. For each property listing, if the 'Floor Number' was found to be greater than the 'Total Floors', the values of the two features were swapped.

	BHK	Rent	Size	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact	Age of Listing (days)	Floor Number	Total Floors
105	1	6000	600	Carpet Area	Pratiraksha Nagar	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner	35	8	5
161	2	10000	450	Carpet Area	Behala	Kolkata	Semi-Furnished	Bachelors/Family	2	Contact Owner	14	2	1

Figure 4. *Rows with greater floor number*

A scatterplot was used to visualize the relationship between the two new features to reveal the patterns and correlations between the two quantitative variables.

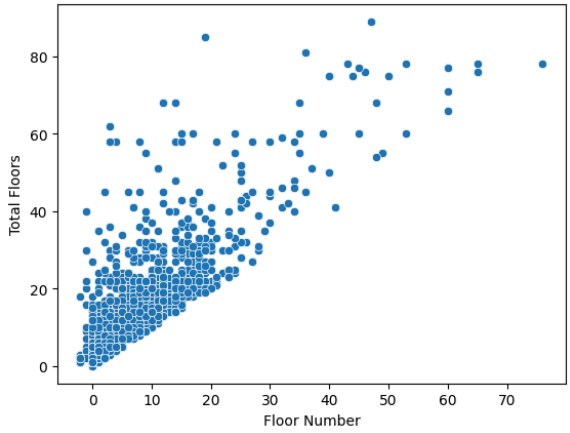

Figure 5. *Relationship between Floor Number and Total Floors*

3.1.3 Feature Selection. It was decided to exclude features, such as 'Point of Contact' and 'Tenant Preferred', because they were

[2] Sameish. "Which Floor is Which?"

hypothesized to have a negligible direct impact on rental prices. They don't influence the rent pricing that much because they are more likely to be a procedural aspect of the renting or leasing. It was also confirmed that including these features didn't affect the model's R^2 score.

On the other hand, the inclusion of 'Area Locality' resulted in a lower R^2 score. This decrease could be due to overfitting or the dilution of more impactful variables' influence. Given a numerous number of unique localities, this feature might have introduced noise rather than useful variance.

3.2 Regularization

In the development of the predictive model, there is a potential for overfitting, especially working with high-dimensional data and capturing non-linear relationships within the dataset. To lighten the risk of this occurrence, Ridge regression was implemented because of its regularization capabilities. This penalizes large coefficients, effectively reducing model complexity.

Among regularization techniques, Ridge regression was selected for its ability to deal with features that affect each other and keeps the model stable by adjusting the importance of features in a balanced way.

The model used the default α value of 1.0. It was determined that this is the optimal α value because of the experimentation with a range, including 0.001, 0.01, 0.1, 1, 10, 100 and 1000. Each was evaluated for its impact on the model's R^2 score with values of 0.001, 0.01, 0.1, 1, and 10 yielding an R^2 score of 0.81, showing a high level of model performance across this range. However, an α value of 100 and beyond resulted in a lower R^2 score. Thus, 1 was chosen to maintain a high R^2 score while effectively regularizing the model.

IV. RESULTS AND ANALYSIS

4.1 Results

4.1.1 Model Performance. The house price prediction model achieved a Coefficient of Determination (R^2) of 0.81, indicating the performance of the model. It shows that it is able to predict up to 81% of the variance explained by the model. The Mean Squared Error (MSE) of the prediction model is 760,954,866.16, which signifies the average squared difference between the observed actual rent prices and the prices predicted by the model.

Table1. *MSE and R^2 scores*

Coefficient of Determination (R^2)	Mean Squared Error (MSE)
0.81	760,954,866.16

4.1.2 Scatterplot Analysis. The scatterplot visualizes the relationship between the actual and predicted rent prices, showing the predictive accuracy of the model. In the plot, it was observed that there is a positive correlation b between the predicted and actual values, with most data points clustered along the line that represents the actual price.

Figure 7. *Actual vs. Predicted Rent Prices Scatterplot*

4.2 Analysis

4.2.1 R^2 Value. The R^2 score of 0.81 indicates that the features selected for the model provided a strong impact for predicting rent prices. It also reflects how the preprocessing steps taken, including feature engineering and the implementation of polynomial features, made the prediction model more accurate.

4.2.2 MSE Value. The relatively high MSE indicates that there is a huge difference between

the model's predictions and the actual prices, especially at the higher end of the rent prices. It could be because of the outliers which were not removed in the model.

V. CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

The model demonstrated a strong capacity to predict rent prices, proven by the high R^2 value obtained. This suggests that the model is capable of accounting a large part of the changes in rental prices, which is an impressive achievement given the complex nature of the housing market. While the Mean Squared Error indicates that there are differences between the predicted and actual rent prices, it shows a meaningful direction for further improvements of the predictive capabilities of the model.

Each feature added into the model impacted the model's performance in predicting real estate market behaviors. The chosen features show what the properties are like and what makes them attractive to pursue. This careful choice of features makes the model reflect the real factors that both owners and consumers care about to determine how much a property is worth, making it a useful tool for figuring out rental prices.

5.2 Recommendations

For future work, these potential improvements are recommended:

(1) Investigate the outliers in the dataset and consider strategies.
(2) Improve the model by incorporating additional features that can impact the rental price and help determine how these features might affect its value in the rental market.
(3) Explore using advanced modeling techniques, which may help improve the model's ability to capture complex patterns within the data.
(4) Add more examples in the datasets, especially from the higher-end of the rent prices range, to ensure that it accurately reflects a border range of rental prices.
(5) Continue adjusting the settings of the model to prevent itself from getting too complicated or too simple. Also, try out different methods to find the best implementation for this model.