Predicting Rent Price in Mumbai

Tanmay Sawaji

---

1. Introduction

1.1 Background

Mumbai is the commercial capital of India and has evolved into a global financial hub. For several decades it has been the home of India's main financial services, and a focus for both infrastructure development and private investment. From being an ancient fishing community and a colonial centre of trade, Mumbai has become South Asia's largest city and home of the world's most prolific film industry. This influx in economy has created a lot of jobs in the city which has resulted in mass immigration from other cities in India and even from outside of India. Since the market is so volatile, employees prefer to rent apartments rather than buying a property.

This study aims to shed light on the contribution of various factors in the pricing of rent in Mumbai city. This study will also help me understand which factors are more prominent in deciding the pricing of rent in Mumbai.

1.2 Interest

This study is primarily for people who are looking for places to rent in Mumbai, as this study will assist them in minimizing the cost while also making sure that they do not compromise in amenities.

2. Data Acquisition and Cleaning

2.1 Data Source

The dataset used in this study is retrieved from kaggle. The dataset is 'Flats for Rent in Mumbai'. The dataset can be found at https://www.kaggle.com/jedipro/flats-for-rent-in-mumbai. The dataset contains 23 columns and 34348 columns. Each row represents a rent pricing in Mumbai. The latest record is of Mid-January, 2020.

2.2 Feature Description and Selection

The Dataset consists of 23 columns but only 8 features are selected as the independant variables and 1 column is selected as the target. The features used are discussed in detailed below in the table. Each row in this data set is a unique property for rent in Mumbai, according to the website 'magicbricks.com' as of mid-January, 2020. The features to be extracted are listed below:

Table 1. Feature list

Column	Description
area	Floor area of the property in sq.ft
bathroom_num	Number of bathrooms available
bedroom_num	Number of bedrooms available
floor_count	Total number of floors in the building
floor_num	Floor on which the property exists
furnishing	Furnishing status; the value can be Unfurnished, Semi-Furnished or Furnished
locality	Locality in which the property is located
user_type	Type of user who posted the ad, ie, Builder, Agent or Owner
price (target)	The rent price of the appartment

2.3 Data Preprocessing/Cleaning

The features ‘furnishing’, ‘locality’ and ‘user_type’ are categorical data and these features need to be encoded as such to be used in the learning model.

The value of feature ‘area’ will be much greater compared to features like ‘bathroom_num’ and ‘bedroom_num’, so these features need to be normalized to remove any bias from the model.

As seen in the dataset, there are a lot of rows with missing data in them. For this model, any row with missing data will be dropped.

The datatype of each feature in the dataset needs to be readjusted. So, the features ‘area’ and ‘price’ are converted to float. The features ‘bathroom_num’, ‘bedroom_num’, ‘floor_num’ and ‘floor_count’ are converted to int as these values should always be a whole number.

Dummy variables are created for all categorical features; these are ‘furnishing’, ‘locality’ and ‘user_type’, which is achieved using ‘One-Hot Encoding’.

3. Exploratory Data Analysis

3.1 Pearson Correlation

There are 4 features which have a continuous value. Pearson Correlation is calculated for these features and the coefficient is listed below:

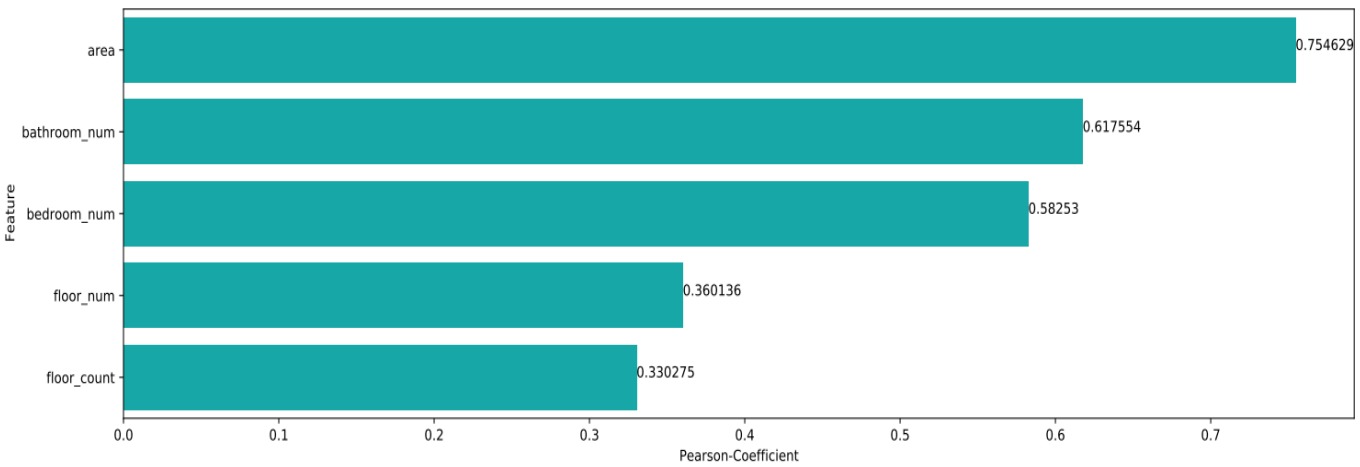
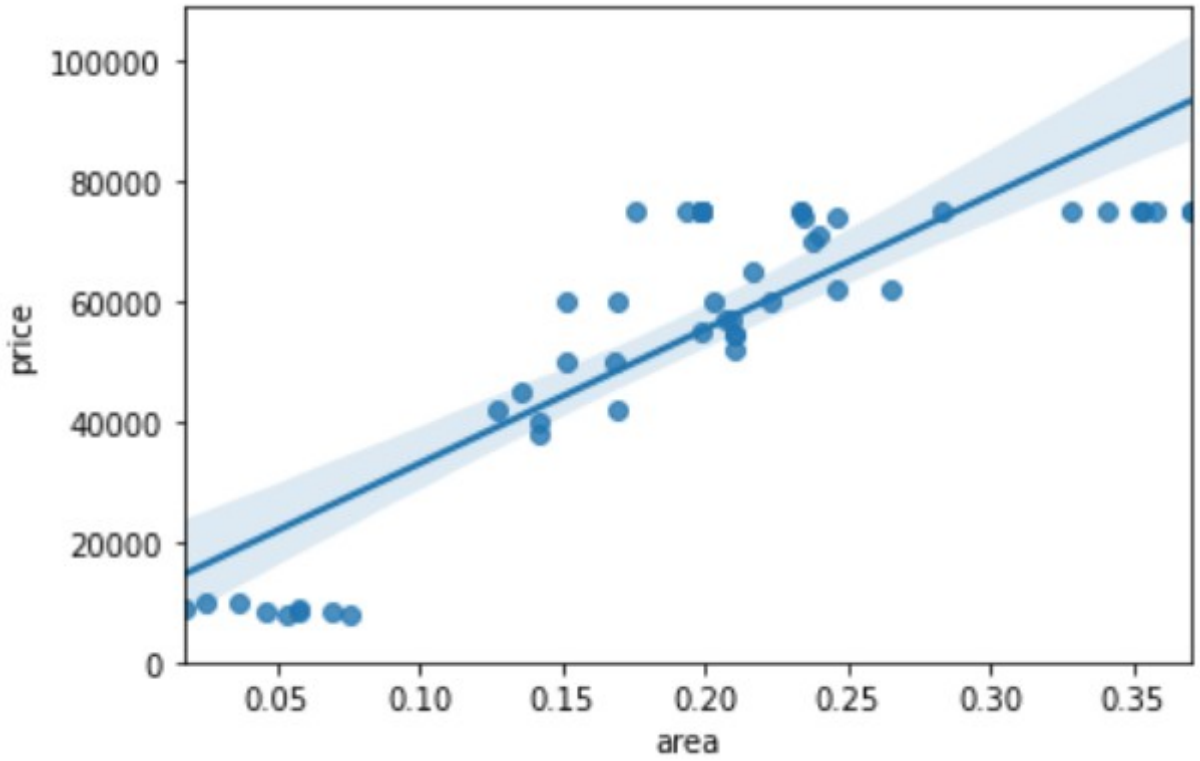


Figure 1. Bar chart showing Pearson Coefficient for various featues

3.2 Relation between Area and Price

The feature ‘area’ represents the carpet area of the entire house. This value is normalized and it lies between 0 to 1. In the graph below we can clearly see that the feature area has a positive linear relation with the target ‘price’. This graph is drawn with only first 50 rows to remove clutter but the relation is the same. This finding is accurate as the size of the house will be directly proportional to its price.

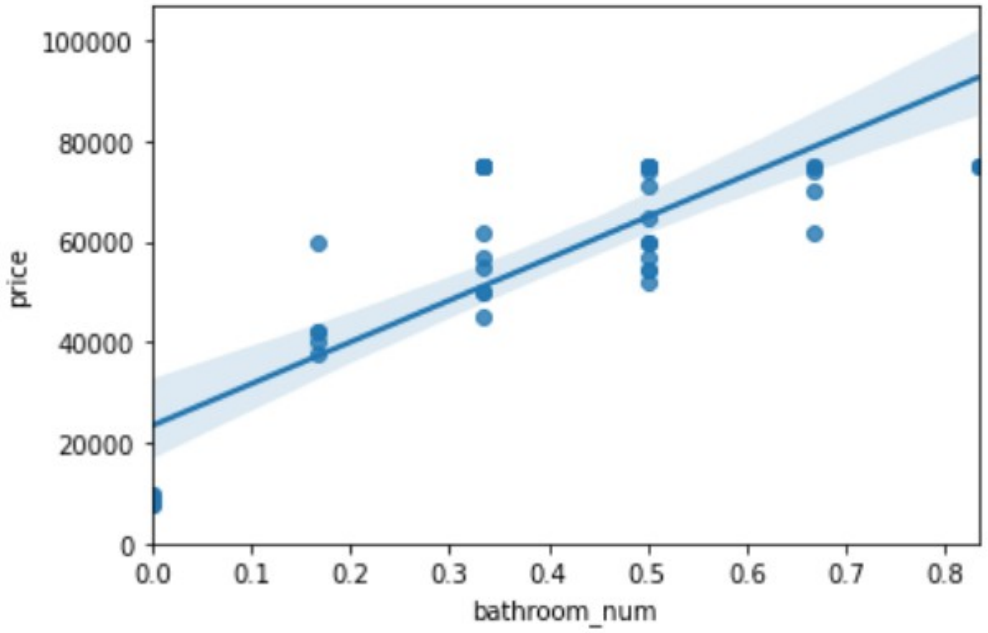
Figure 2. Regression Plot between area and price



3.3 Relation between Bathroom_num and Price

The feature ‘bathroom_num’ represents the number of bathrooms present in the house. This value was initially a whole number but it was normalized and now its value lies between 0 and 1. In the graph below we can clearly see that this feature has a positive linear relation with the target ‘price’. This graph contains only the first 50 rows of the dataset to remove clutter in the graph but this does not change the observations made. This result was expected as increase in the number of bathrooms in a house would, in general, increase the price of the house.

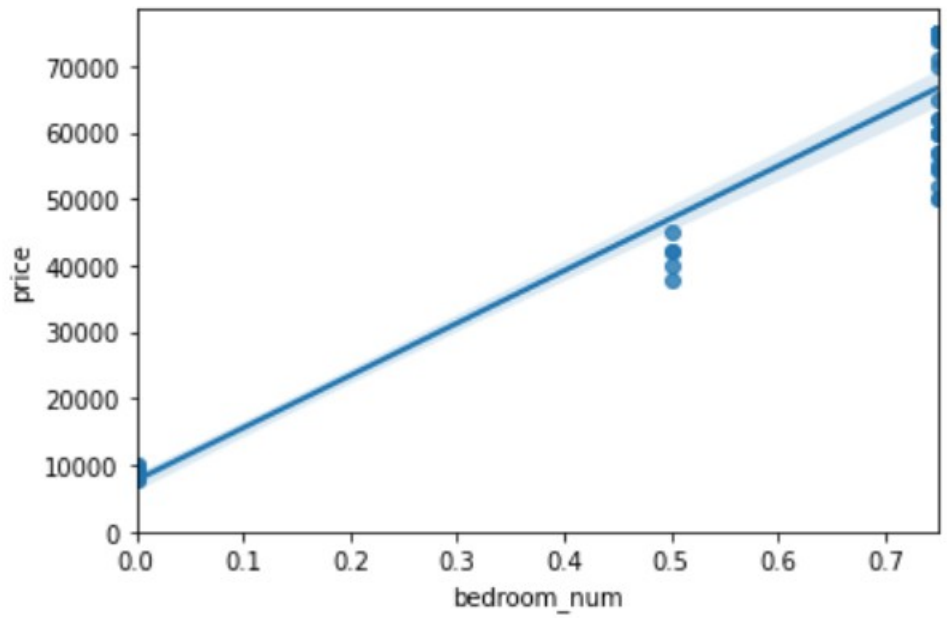
Figure 3. Regression Plot between bathroom_num and price



3.4 Relation between Bedroom_num and Price

The feature ‘bedroom_num’ represents the number of bedrooms present in the house. This value was initially a whole number but it was normalized and now its value lies between 0 and 1. In the graph below we can clearly see that this feature has a positive linear relation with the target ‘price’. This graph contains only the first 50 rows of the dataset to remove clutter in the graph but this does not change the observations made.

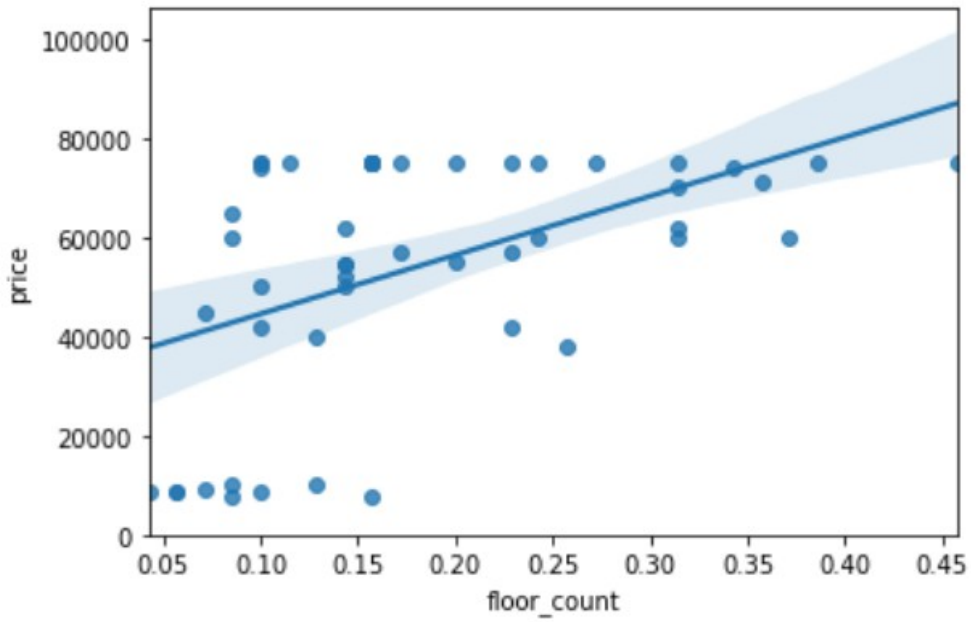
Figure 4. Regression Plot between bedroom_num and price



3.5 Relation between Floor_count and Price

The feature ‘floor_count’ represents the number of floors of the building that the house belongs to. This value is normalized and lies between 0 and 1. In the graph below we can see that this feature has a positive linear relation with the target ‘price’.

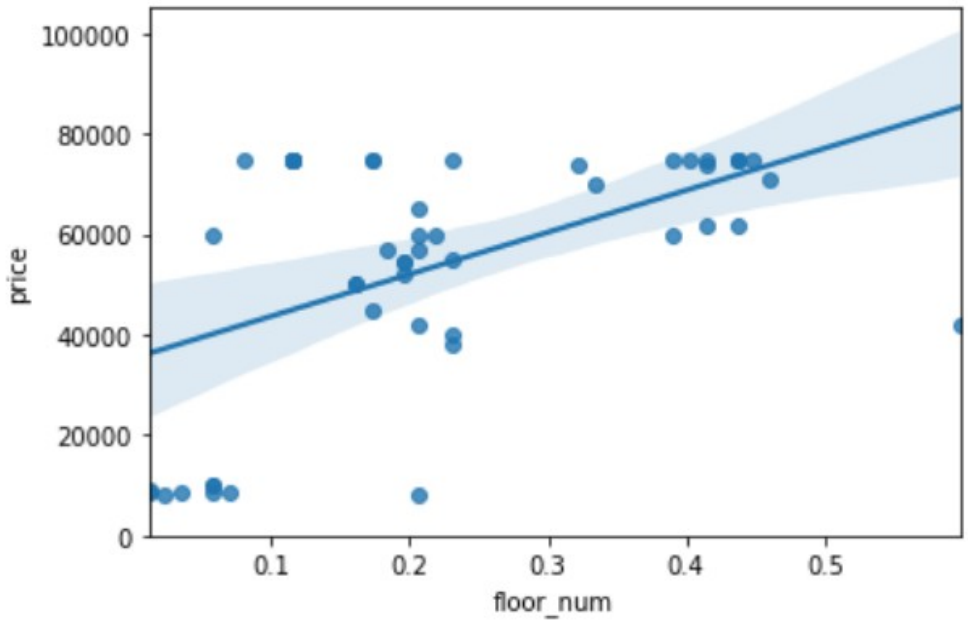
Figure 5. Regression plot between floor_count and price



3.6 Relation between Floor_num and Price

The feature ‘floor_num’ represents the floor number of the house. This value is normalized and lies between 0 and 1. In the graph below we can see that this feature has a positive but weak linear relation with the target ‘price’.

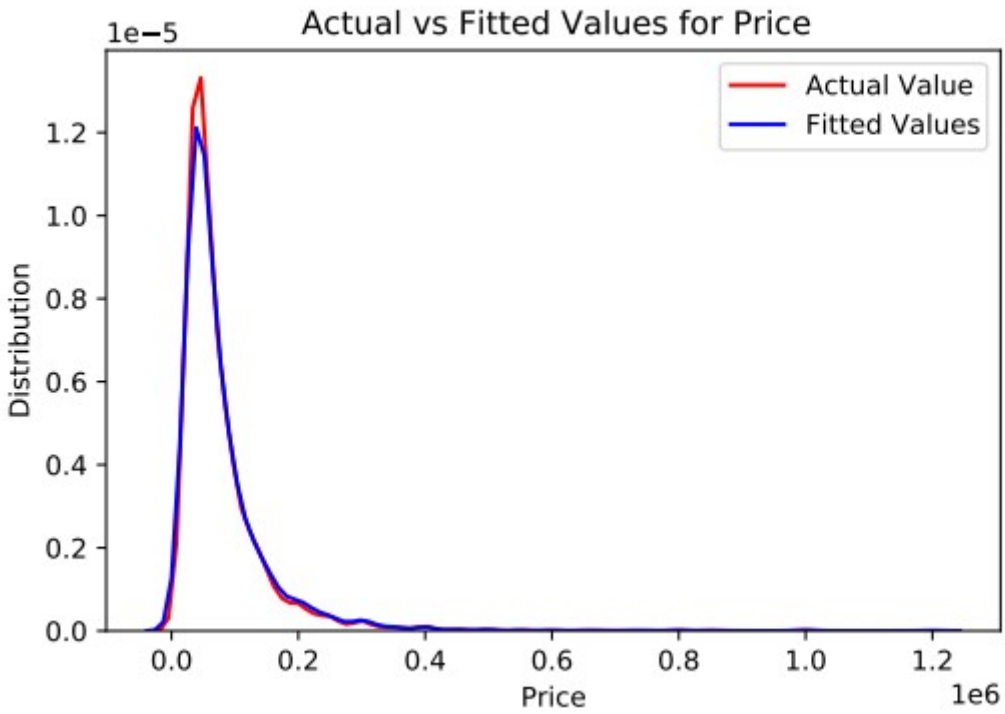
Figure 6. Regression plot between floor_num and price



4. Model Deployement and Evaluation

The dataset has multiple features that are linearly correlated to the target variable as seen in the previous section. This means that a Multiple Linear Regression model will be best suited for the dataset selected. A linear regression model is created using the Scikit-learn library and the distribution plot of the result is shown below. This graph shows that the model fits the data very well and no adjustments are requried. The data split chosen was 60% training and 40% testing data.

Figure 7 Distribution Plot of the Regression Model



5. Result

In this study, the correlation of each continuous feature to the target price was calculated. This shows that the features selected and the preprocessing of these features improved the final fit of the multiple linear regression model selected. This correlation can be seen in the table below.

Feature	Pearson Coefficient	P-Value
Area	0.7546293094784234	0.0
Bedroom_num	0.5825303173512928	0.0
Bathroom_num	0.6175540260345069	0.0
Floor_num	0.36013593950495626	7.081348519710123e-305
Floor_count	0.33027522488100225	7.398827032795715e-254

Table 2. Pearson Coefficient and P-value of all features

The intercept and coefficients of the model are calcualted and shown in the table below to show the impact of all features in the model.

Parameters	Value
Intercept	2.910383e-11
area	8.090929e-10
bathroom_num	-2.169513e-10
bedroom_num	-1.166648e-11
floor_num	1.438985e-11
floor_count	3.479912e-12
Furnished	1.503536e-12
Semi-Furnished	-1.291026e-12
Unfurnished	-2.125097e-13
locality_Andheri East	-1.378207e-12
locality_Andheri West	1.255561e-12
locality_Bandra West	2.586060e-11
locality_Bhandup West	-3.915250e-12
locality_Chembur	-3.674419e-12
locality_Goregaon East	-5.862841e-12
locality_Mulund West	-8.571469e-12
locality_Parel	-3.906141e-13
locality_Powai	1.568108e-12
locality_Thakur Village, Kandivali East	-3.358902e-12
locality_Worli	-1.532565e-12
Agent	-9.290954e-12
Builder	1.691434e-11
Owner	-7.623389e-12

Table 3. Model Intercept and Coefficients

6. Conclusion

A model accurately predicting the rent price of houses in the city of Mumbai is successfully created and tested. This model shows that the feature that affects the price of a house is the carpet area, followed by the bedroom number and then the bathroom number. The locality also significantly affects the price of the house.