Real Estate Rent Prediction Model Using Linear Regression

CS ELEC 2C (Machine Learning) - Lab Exercise # 1

Dela Paz, Angelo Daniel A.

I. Introduction

Housing in India has grown significantly in recent decades. This increase is due to a significant increase in income in the country recently, with the per capita income across the country increasing by around 50% from 2015 to 2019 (*India: Per Capita Income 2023 | Statista*, 2023). Housing in the country may range from palaces to modern apartments in urban areas.

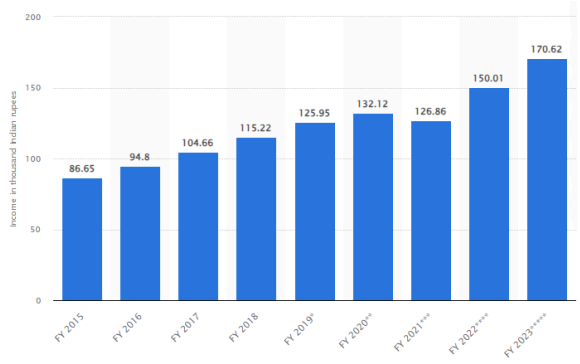


Figure 1.1 Per capita income across India from 2015 to 2023

This literature aims to solve the problem of predicting the rental prices of various housing in India based on a given data set of rental prices for numerous real estate properties. A linear regression model will be used to accurately predict the prices of the properties based on several factors, including the number of floors and rooms, the property's size, and the property's location.

Since a linear regression model will be developed for this problem, several aspects regarding the properties of a linear regression model should be considered with the features used. Firstly, a linear regression model requires the use of continuous variables to map the relationships between features. The need for continuous variables presents an issue in the data set since some features, such as city, preferred tenant, and area type, are categorical rather than ordinal. The processing of these features is essential during the preprocessing stage to make them usable for the model. Another consideration when using linear regression models is outliers in the data set. Linear regression is very sensitive to outliers and can affect the final accuracy of the predictions made by the model.

It is important to note that the model to be developed should not be able to predict with perfect accuracy due to the risk of overfitting the training data, which would hamper the accuracy of the test predictions. The reasonable coefficient for the model to be developed would be above 65% for the coefficient of determination.

II. Methodology

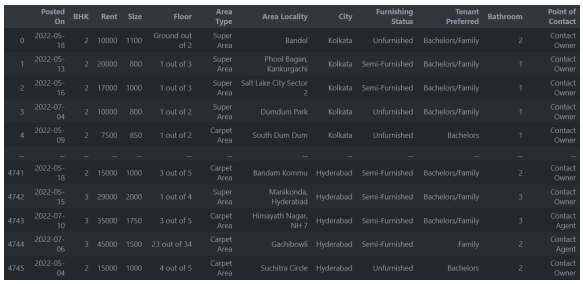
The model will be utilizing several libraries to complete its development fully. First, Pandas will be used to manipulate the data set itself. It allows the use of the DataFrame structure, which will be used to store the .csv data set. Pandas also has excellent compatibility with other libraries like NumPy and Scikit-learn. NumPy will be used for its scientific computation capabilities, specifically on array and matrix operations. The linear regression will be called from the Scikit-learn, as the Library is primarily used for machine learning and predictive analysis. Lastly, Matplotlib and Seaborn will be used for their data visualization capabilities in showing the relationships between the independent and dependent variables.

The data set used in this model is a .csv file of 4,747 entries that describes the rent of properties in India. The data set can use ten columns for training the linear regression model. The possible features are the following:

- **BHK**.

- **Size**
- **Floor**
- **Area Type**
- **Area Locality**
- **City**
- **Furnishing Status**
- **Tenant Preferred**
- **Bathroom**
- **Point of Contact**

The rent column of the data set will be the dependent variable in the model since the model will predict the possible price of a property. Before modeling, the data set will undergo preprocessing, which includes data cleaning, feature engineering, encoding, splitting, standardization, and regularization techniques.



	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

Figure 2.1 Original Data Set

In preprocessing, all except two of the features (Posted On, Area Locality) mentioned above will be used in training the model. The decision to remove the "Area Locality" feature stems from the diverseness of the values, which would be difficult to encode without adding many columns during encoding.

Outlier detection using the Interquartile range (IQR) was first performed on some features to maintain consistency during the training and testing phase (Bhat, 2023). To effectively ingest all features into the model, all features with string values will be converted into a numerical value for ingestion into the model. To convert string data into a numerical value, one-hot encoding will be performed on categorical features to convert each categorial value to a new column and assign a binary value if they apply to each row (*Data Science in 5 Minutes: What Is One Hot Encoding?*, n.d.).

The data set was then split using an 80/20 test/train split. This split is the most common train/test split for machine learning projects as it provides enough data for sufficient training while having enough for a reliable test sample. Lastly, standardization was performed on the training set using StandardScaler and Polynomial features from the Sckit-learn library.

III. Experiments

3.1 Data Cleaning

3.1.1 Feature Selection

As mentioned in the methodology section of the literature, eight out of the ten features were chosen to develop the linear regression model. The "Posted On" feature was not included in the feature selection because it is inferred to be insignificant in the overall rent value of the property. Additionally, it adds another layer of difficulty in converting dates into a meaningful feature for the model. Area Locality is another feature that was omitted from the training features due to its difficulty in converting many categories into meaningful numerical data for the model.

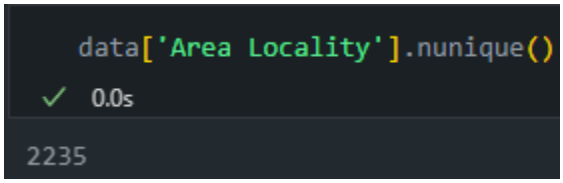


Figure 3.1.1 Count of Unique "Area Locality" Values

Other features, such as City and Preferred Tenant, were considered for dropping; however, after encoding, they contributed to the increase of the determination coefficient, albeit slightly.

3.1.2 Detection and Removal of Outliers

Outliers in each feature of the data set were detected using the Interquartile range of each feature.

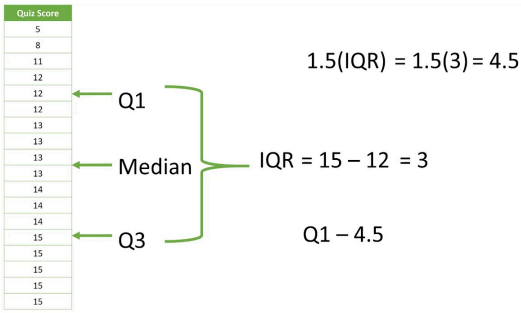


Figure 3.1.2 Identification of Outliers through IQR

As seen in the figure above, outliers are found outside the upper and lower bounds of the 1st and 3rd quartiles of the data set. The same procedure was also performed for the data set's Rent, Bathroom, and Size features. These features were chosen through distribution plot analysis, which showed significant outliers in their graphs. The model drops these rows to address these outliers to make the distribution more standardized.

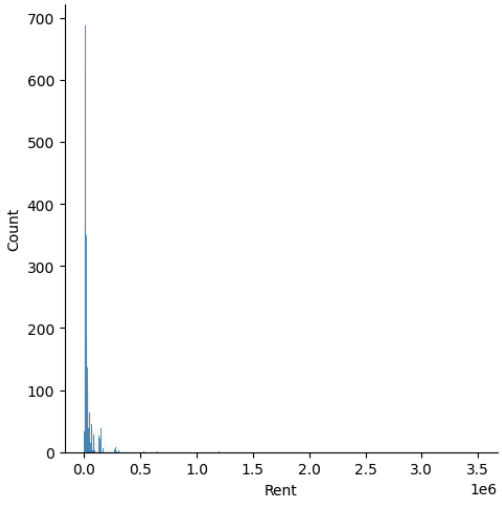


Figure 3.1.3 Distribution of Rent Counts

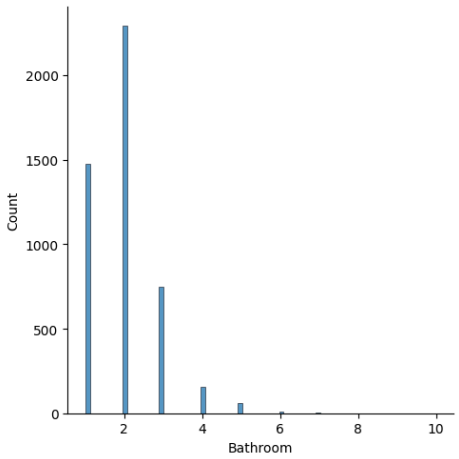


Figure 3.1.4 Distribution of Bathroom Counts

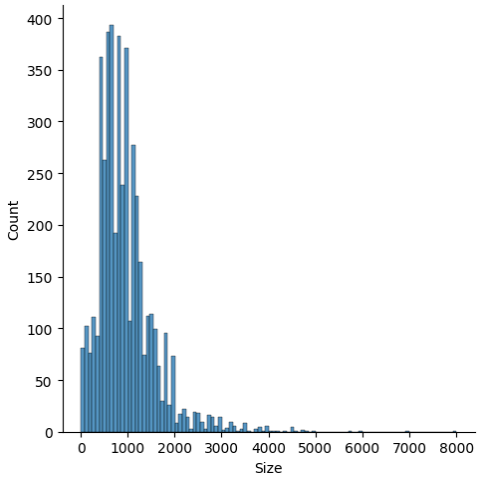


Figure 3.1.5 Distribution of Size Counts

3.2 Conversion of Features with String Data to Numerical Data

3.2.1 Floor Feature Splitting

One of the problematic features that needed to be addressed for the model is the "Floor" feature. The feature values have the following format: "X out of Y," wherein X is either a numerical or string value and Y is a numerical value. This format for values would not be able to be ingested into the model due to the string values.

Two possible solutions were considered to address this problem. One solution would be to split the floors into two columns, namely "Current Floor" and "Total Floors". The other solution would be

to combine the numerical values by dividing them into one value without the string characters. It was decided to proceed with the first approach since it results in more consistent values because the second approach would lead to more values due to the division step.

It is noted that for the X value, it is possible to have strings as their value, such as "Ground," "Upper Basement," and "Lower Basement." To address them, they will be assigned a value of 1, 0, and -1, respectively. Both newly created feature values are then increased by two to scale their values appropriately.

3.2.2 One-Hot Encoding

Other categorical values, such as city, furnishing status, point of contact, etc., were converted to numerical values using one-hot encoding. One-hot encoding is a technique used to convert categorical values into numeric values for a machine learning model (*One Hot Encoding in Machine Learning*, 2023).

Fruit	Categorical value of fruit	Price
apple	1	5
mango	2	10
apple	1	15
orange	3	20

The output after applying one-hot encoding on the data is given as follows,

apple	mango	orange	price
1	0	0	5
0	1	0	10
1	0	0	15
0	0	1	20

Figure 3.2.1 Simple one-hot encoding example

After applying the technique to the applicable features, the resulting data frame consisted of nineteen features from the original eleven features.

3.3 Standardization

Standardization is a common requirement for most machine learning models. A non-standardized dataset can behave inappropriately if it does not look like a standard normal distribution. The project used two standardization classes from Scikit-Learn to apply standardization to the dataset, namely StandardScaler and PolynomialFeatures.

The StandardScaler is a preprocessing utility class in the Scikit-learn library that standardizes the features by removing the mean and scaling to unit variance in the data frame. This class subtracts the data set's mean from each value and divides the result by the standard deviation.

3.4 PolynomialFeatures

This utility class generates a new feature matrix consisting of all polynomial combinations of the features with a degree less than what is specified. This class was used to capture the more complex relationships of the data set.

Although this class is commonly used for polynomial regression, since it models a non-linear relationship, it can be considered a linear model as the regression function used is linear in terms of its parameters.

3.5 Regularization

The primary use of regularization is to prevent overfitting data by adding a penalty variable to the loss function of large weights. The model used two kinds of regularization in Ridge and Lasso that yielded similar results, which will be discussed in the results section.

IV. Results and Analysis

Figure 4.1 Scatterplot Analysis of Actual vs. Predicted Prices

In the figure above, the blue dots represent the predicted price of the property made by the regression model, while the green dots represent the actual price of the property. It can be observed that predicted prices are much closer to their actual values at lower rent values and increase the higher the value of the rent becomes. This indicates that the model is more accurate in predicting the prices of low-value properties than higher-valued ones. This can be due to the low count of higher-priced properties where values vary greatly. Another observation regarding the figure above is the error of the predicted values places them below the actual price. This indicates that the model tends to underestimate the value of the property. This may be due to the regularization techniques wherein the coefficients were pushed closer to zero.

	Value
Mean Squared Error	44332760.02
Coefficient of Determination	0.76

Figure 4.2 MSE and R^2 Score of Test Set

In Figure 4.2, the coefficient of determination and the mean squared error (MSE) of the test data set can be observed. The MSE score of 44,332,760 is one of the biggest factors to improve upon in this model. Like the coefficient of determination, the preprocessing techniques greatly reduced the MSE. As stated in the introduction, the model is not designed to predict all property prices perfectly; however, a score of 0.76 is a substantial enough score to accurately predict the prices of real estate properties most of the time. An initial overview of the techniques used in preprocessing indicated that the PolynomialFeatures and One-hot encoding techniques were the biggest contributing factors to the increase in the final R squared score. This may be because of the increased total features the model ingested before modeling.

V. Conclusions & Recommendations

Based on the analysis of the model's accuracy on the test data set, it can be concluded that the linear regression model can accurately predict to a reliable degree. A coefficient of determination of 0.76 means that the model only misses its predictions less than 25% of the time.

Further research and experimentation are recommended to see if the model can be improved with better accuracy. Some suggestions are including the 'Area Locality' feature through other encoding techniques, such as binary or label encoding. There is also the possibility of using other models rather than a linear regression model, which can capture more complex relationships between the features and property rent. Additionally, the underestimation of higher-valued properties should also be further explored to identify root causes and possible solutions.

VI. References

Bhat, H. (2023, August 18). *How to Detect Outliers in Machine Learning? (With Examples).* AlmaBetter. https://www.almabetter.com/bytes/articles/outlie

r-detection-methods-and-techniques-in-machine-learning-with-examples

Data Science in 5 Minutes: What is One Hot Encoding? (n.d.). Educative. https://www.educative.io/blog/one-hot-encoding

India: per capita income 2023 | Statista. (2023, August 23). Statista. https://www.statista.com/statistics/802122/india-net-national-income-per-capita/

One Hot Encoding in Machine Learning. (2023, April 18). GeeksforGeeks. https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python/