Analyzing Rental Rates in India with Linear Regression

Advanced Intelligent Systems (CS-ELEC2C) Laboratory Exercise # 1

Abrigo, Nathanael Chris O.

I. Introduction

Housing in India consists of different settings, from palaces of erstwhile maharajas to modern apartment buildings. It also consists of other areas, such as a bustling city and villages away that are isolated. As the demand in India's housing sector continues to rise tremendously, rental rates also increase. Analyzing rental rates will require multiple details to ensure its validity and consistency. According to Kelly (2022), pricing factors for renting properties include the overall situation of the property itself, such as location, size, condition, and many more.

Using linear regression can address the problem of multiple factors affecting the solution. According to Ghosh (2022), Linear Regression is used to predict or forecast a value based on dependent variable/s associated with it. Ghosh (2022) also stated that businesses can use linear regressions to analyze future trends and make estimates or forecasts. Linear regression can be utilized to provide the output and rental rates using the inputs, which are the factors for the rental rates.

However, Linear Regression will require tweaking inputs that are not numerical, such as the property's location. Also, this method can only be partially accurate as this is a regression approach. Although given the limits that Linear Regression imposes, it is still reliable to provide predictions that are acceptable to a reasonable degree.

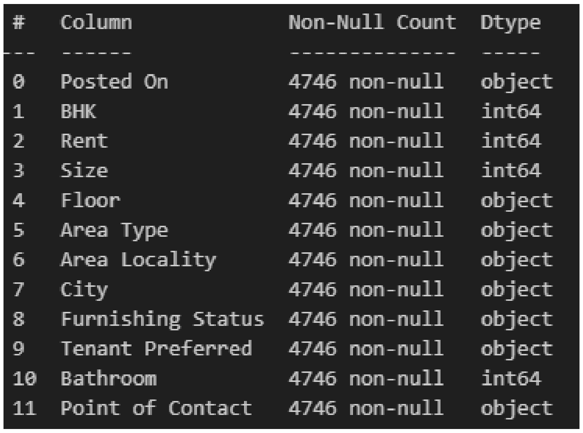
II. Methodology

This section will discuss the steps taken for the program to process the information to provide the desired output.

1. Setup

This step will involve installing and initializing the necessary libraries such as numpy, pandas, matplotlib.pyplot, and seaborn. This step will also include importing the dataset for manipulation.

2. Analysis



```
#   Column            Non-Null Count  Dtype
--- ------            --------------  -----
0   Posted On         4746 non-null   object
1   BHK               4746 non-null   int64
2   Rent              4746 non-null   int64
3   Size              4746 non-null   int64
4   Floor             4746 non-null   object
5   Area Type         4746 non-null   object
6   Area Locality     4746 non-null   object
7   City              4746 non-null   object
8   Furnishing Status 4746 non-null   object
9   Tenant Preferred  4746 non-null   object
10  Bathroom          4746 non-null   int64
11  Point of Contact  4746 non-null   object
```

Figure 2.1: Dataset columns

This step will involve understanding the dataset. The dataset provided for this study is a CSV file comprising 11 features. These are the date of posting, bedroom, hall, and kitchen (BHK) count, rent cost, size in square feet, floor location, area type, area locality, city, furnishing status, tenant preferred, bathroom count, and point of contact.

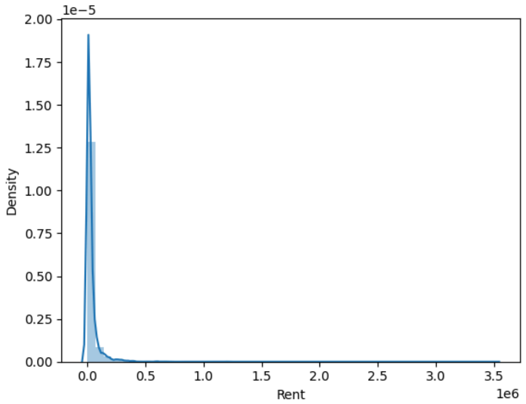
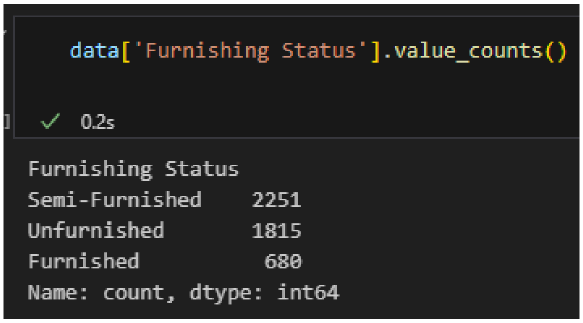


Figure 2.2: sns.distplot(data['Rent']) graph

This step will also involve using matplotlib.pyplot, and seaborn libraries to visualize the data. For instance, sns.distplot(data['Rent']), this code snippet showed that there are data that are considered outliers that can be removed.



```
data['Furnishing Status'].value_counts()
```

✓ 0.2s

```
Furnishing Status
Semi-Furnished    2251
Unfurnished       1815
Furnished          680
Name: count, dtype: int64
```

Figure 2.3: Using value_counts()

This step will also utilize methods such as .value_counts() and .unique() to check other features. For instance, data['Floor'].value_counts() showed that data needs to be fixed since most are inputs like '1 out of 3,' but some are 'Ground out of 3'.

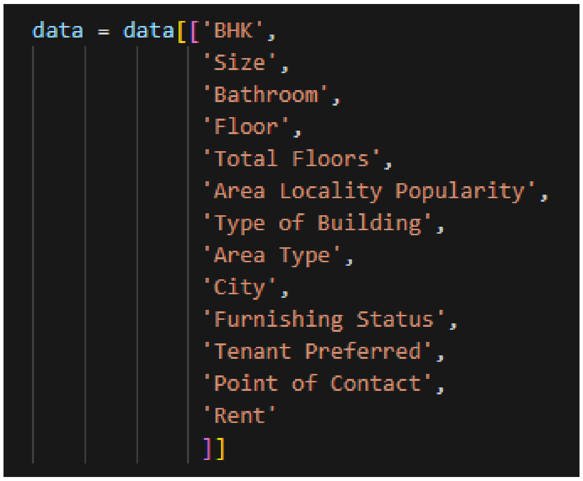
3. Cleaning of Data

This step will comprise cleaning and manipulating data during the study. As mentioned earlier, at data['Floor'], there are inputs like '1 out of 3', but some are 'Ground out of 3'. There are also values like 'Upper Basement out of 3' and 'Lower Basement out of 2'. This step will transform this data into the following equivalent value: 'Ground' into 1, then for 'Upper Basement' and 'Lower Basement' inputs will be dropped since both inputs will make the 'out of' statement vague since it may or may not include in the count which can reduce accuracy.

Next, also in the 'Floor' column, there are inputs that the left side is greater than the right side. For instance, '8 out of 5' and '2 out of 1'. This input is transformed into '5 out of 8' and '1 out of 2' to be standardized in all the 'Floor' columns.

4. Preprocessing for Model

This step will prepare the features for the model to use. Out of the 11 features, the date of posting is not included in the preprocessing for the model.



```
data = data[['BHK',
             'Size',
             'Bathroom',
             'Floor',
             'Total Floors',
             'Area Locality Popularity',
             'Type of Building',
             'Area Type',
             'City',
             'Furnishing Status',
             'Tenant Preferred',
             'Point of Contact',
             'Rent'
             ]]
```

Figure 2.4: Encoding features

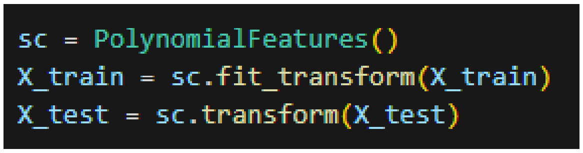
For the features 'Floor' and 'Area Locality,' these features were not directly used but were used as a basis to derive some information for 'Floor,' this data can derive the 'Total Floors' of the building, 'Floor Rented' or which floor is being rented, and 'Type of Building.' According to Sarac (2019), Low-rise buildings are buildings with four floors or under. Mid-rise buildings are buildings that have between 5 to 12 floors. High-rise buildings are buildings that have 13 floors or above. Skyscrapers are buildings with over 40 floors.

For 'Area Locality,' this input was used to derive 'Area Locality Popularity.' For instance, 'Murugeshpalya, Airport Road' appeared fourteen times in the whole data set, 'Area Locality Popularity' will record the number '14' since this number signifies how often it is seen in the dataset as location popularity is also a factor in rent prices.

As for other string inputs like for 'BHK,' 'Type of Building,' 'Area Type,' 'City,' 'Furnishing Status,' 'Tenant Preferred,' and 'Point of Contact,' the program used a method named one_hot_encode(data, column) to make it boolean value for the model to process the data.

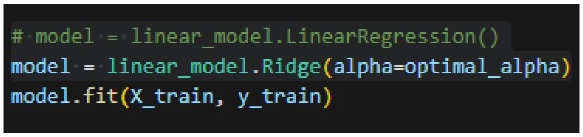
Other number value inputs like 'Size,' 'Bathroom,' 'Floor Rented,' 'Total Floors,' 'Area Locality Popularity,' and 'Rent' have no problems since the model can process the data.

5. Evaluation



```
sc = PolynomialFeatures()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

Figure 2.5: using polynomial features



```
# model = linear_model.LinearRegression()
model = linear_model.Ridge(alpha=optimal_alpha)
model.fit(X_train, y_train)
```

Figure 2.6: using Ridge instead of Linear

This will utilize the Ridge regression model with polynomial features to process the data and find out the mean squared error and coefficient of determination to train and test the program.

III. Experiments

This section will discuss the attempts made and explain why the attempt is made the way it is.

1st Attempt

After studying the initial program given, it is observed that some features are not included in the encoding procedure. That is why it was decided to add the features 'Size,' 'Area Type,' and 'City' to the encoding procedure since these values also affect the rent price of the property.

2nd Attempt

After making the 1st Attempt, there has been a deliberation regarding how the program can utilize the data from ‘Floor.’ It was later decided to create a new column, ‘Total Floor,’ to record the total floor of the property since knowing how high a property is is also a factor when checking the place to rent.

3rd Attempt

After making the 2nd Attempt, it was realized that the ‘Floor’ is necessary to have in the encoding procedure since floors are factors in the renting price. It is decided to make the ‘Floor’ value into its equivalent float value so the program can recognize it. For instance, ‘1 out of 4’ will become 0.25.

4th Attempt

After the 3rd Attempt, there has been an idea regarding how buildings are categorized, such as how many floors are considered high-rise. This led to finding out that there are specific floors to consider a building as a low-rise, mid-rise, high-rise, or skyscraper, which was applied in the program as a new column ‘Type of Building’ based on ‘Total Floors.’ The 3rd attempt details are also removed since it increases the mean squared error.

5th Attempt

After making the 4th Attempt, there has been a dilemma for the ‘Upper Basement’ and ‘Lower Basement’ since there may be ambiguity in the data, like if the basement floors are included in the count of total floors. This led to creating a new column, ‘Basement,’ on the encoding procedure to differentiate which are basements and which are not. During this attempt, basement levels were not dropped, which was previously mentioned in the cleaning of data phase.

6th Attempt

After the 5th Attempt, there was a moment of confusion as to what more data could be manipulated to further increase the coefficient of determination. A tip was mentioned by a friend, which is to remove outliers. After realizing what outliers mean, removing them was implemented in the program immediately. In this attempt, 'Tenant Preferred' and 'Point of Contact' were added to the encoding procedure since these also slightly affect the predicted rental price. 'Floor Rented' was also added to represent the property's specific floor.

7th Attempt

After doing the 6th Attempt, there have been realizations regarding the basement floors. There have been deliberations on whether basement floors are considered in the floor count. And when transforming the values of ‘Upper Basement’ into ‘-1’ and ‘Lower Basement’ into ‘-2’ is accurate. This led to the idea of dropping the row that entertains basement floors as it creates confusion in the data. This attempt also removed the ‘Basement’ column in the encoding procedure as it is not applicable already.

8th Attempt

After making the 7th Attempt, there are realizations if my 3rd Attempt is returned to the program since the possible ambiguity, which is the basement floors, is removed in the previous attempt. This led to returning the ‘Floor’ that is transformed to its decimal value, and this prompted the removal of ‘Floor Rented’ on the encoding procedure as ‘Floor’ itself already represents which floor property is being rented.

9th Attempt

After making the 8th Attempt, there have been ideas regarding tweaking the program's model. This led to switching to PolynomialFeatures() instead of StandardScalar() since PolynomialFeatures() can capture more complex relationships between the features and the target variable. Also, LinearRegression() was changed to use Ridge(), a Regression algorithm. Ridge Regression is utilized since it adds a regularization term to the linear regression objective function, which prevents overfitting by reducing the variance of the model.

IV. Results and Analysis

This section will discuss the results obtained from the experiments and some insights as to how it arrived at that conclusion.

Initial Model Results

```
Mean squared error:
2693550407.36
Coefficient of determination:
0.32
```

1st Attempt Results

```
Mean squared error:
1905735826.05
Coefficient of determination:
0.52
```

The 1st Attempt Results show how significant the features 'Size,' 'Area Type,' and 'City' are in the encoding procedure. This means that these features are crucial to the rent pricing.

2nd Attempt Results

```
Mean squared error:
1893363234.27
Coefficient of determination:
0.52
```

The 2nd Attempt Results show a slight improvement in the MSE. This means that creating the ‘Total Floor’ column has a positive effect, although not significantly.

3rd Attempt Results

```
Mean squared error:
1913848680.58
Coefficient of determination:
0.52
```

The 3rd Attempt Results show a slight increase in the MSE. This means that making the ‘Floor’ value decimal does not improve the prediction. This may also mean that the ‘Floor’ still has some uncertain values, affecting the program negatively.

4th Attempt Results

```
Mean squared error:
1878403372.33
Coefficient of determination:
0.53
```

The 4th Attempt Results show a slight improvement in the MSE and coefficient of determination. Adding the ‘Type of Building’ helped the program further understand the relationships between features.

5th Attempt Results

```
Mean squared error:
1878398734.83
Coefficient of determination:
0.53
```

The 5th Attempt Results show a slight improvement in the MSE. This means adding ‘Basement’ helped a little in separating the floors that are not basement levels.

6th Attempt Results

```
Mean squared error: 57299293.93
Coefficient of determination:
0.70
```

The 6th Attempt Results show a considerable improvement in the MSE and coefficient of determination. This means that removing outliers is crucial since it removes values that are very far from the average, which negatively affects the prediction of the program.

7th Attempt Results

```
Mean squared error: 51750588.36
Coefficient of determination:
0.71
```

The 7th Attempt Results show a slight improvement in the MSE and coefficient of determination. This means that the removal of basement floors, which provides confusion, is an option to reduce noise in the prediction.

8th Attempt Results

```
Mean squared error: 51748467.56
Coefficient of determination:
0.71
```

The 8th Attempt Results show a slight improvement in the MSE. This means that the return of the 'Floor' to decimal equivalent was now better than before since the possible noise in that feature was dropped.

9th Attempt

```
Mean squared error: 45916504.99
Coefficient of determination:
0.74
```

The 9th Attempt Results show a good improvement in the MSE and coefficient of determination. This means that the change in the model is also essential, which also means that the current model being used (Ridge Regression) is more suitable than the standard Linear Regression model.

V. Conclusions & Recommendations

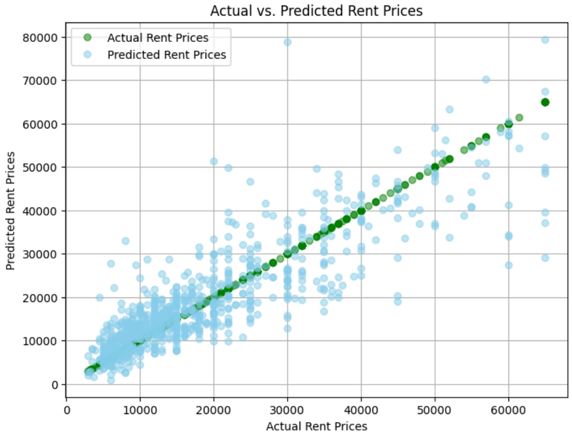


Figure 5.1: Actual vs Predicted Rent Prices

With the final Mean Square Error (MSE) of 45916504.99 and final Coefficient of Determination of 0.74, while the coefficient of determination is still far from the perfect prediction of 1, it is still pretty good. However, the MSE of the model is still very high, which means there are still possible improvements that can be made to the model. Having an elevated MSE also means that the program tends to estimate the rent price off the charts.

During the experimentation phase, multiple trials and errors and realizations about which features in the data set are crucial to analyzing the rent price. 'Size' is one of the most critical features that contributes to the increase of the coefficient of determination of the program. Features like 'Floor' are also essential, but data cleaning and manipulation are necessary to utilize the data in that column. String features like 'City' require the one_hot_encode method to help the program process the value.

Although the standard Linear Regression model with Standard Scalar works fine, the Ridge Regression with Polynomial Features helped the program improve its accuracy in identifying the MSE and coefficient of determination. This means that using a more suitable model can further improve the program's accuracy.

Based on the findings that are discovered during the course of the study, it is recommended that a suitable model that can provide better accuracy is found. Ridge Regression with Polynomial Features proved better in this study than the standard Linear Regression model with Standard Scalar. However, there may also be more models that can outdo Ridge Regression that still need to be discovered.

Another recommendation is to provide further features that can factor into the renting price of a property. For instance, accessibility to the place could include public transportation, grocery stores, schools, and more. These possible features may provide better accuracy to the program, given that the data is processed thoroughly to ensure its integrity.

VI. References

Ghosh, B. (2022, June 24). *The Magic of Linear Regression Model*. Linked In. https://www.linkedin.com/pulse/magic-linear-regression-model-bhagyashree-ghosh/

Kelly, H. (2022, July 28). *12 Rental Pricing Factors 2022: How to Price Your Rental Propert*y. Belong Home. https://belonghome.com/blog/rental-pricing-factors

Sarac, F. (2019, November 20). *High-Rise Apartment Buildings and Mid-Rise Buildings Are Overshadowing Low-Rises for the First Time in 3 Decades*. RentCafe. https://www.rentcafe.com/blog/apartmentliving/high-mid-rise-residential-buildings-overshadowing-low-rise/#:~:text=High%2Drise%20buildings%20are%20defined,of%20the%20high%2Drise%20category