Rent Price Prediction Model Using Linear Regression

Advanced Intelligent Systems (CS-ELEC1A) Laboratory Activity

Mala-ay, Amiel Christian E.

I. INTRODUCTION

Regression problems can be encountered throughout various fields of study. Various problems can be addressed with linear regression, so long as the variables are continuous or can be manipulated to be so. The use of linear regression can help in modeling the relationship between different variables and make predictions based on this relationship. Problems addressed through linear regression include stock price prediction, consumer behavior, and sales forecasting. Another such problem is housing price prediction, where rent can be viewed as a variable dependent on factors such as a rental's size, its location, amenities, etc.

In the problem of predicting rent prices with linear regression, however, there are some caveats. For one, linear regression needs continuous variables, so there are challenges to face when dealing with non-quantitative data, of which there is a lot in the problem of rent price prediction. Features like a rental's location, furnishing status, area type, etc. cannot be used outright by a linear regression model. As such, there is a need to make manipulations on such data.

Moreover, a linear regression model cannot be expected to make predictions with total accuracy. It is incredibly rare for real-world data to follow a line perfectly, so linear regression cannot always be accurate in its predictions. Making a linear regression model fit exactly with the data it is trained on would, in fact, hamper its performance and reduce its ability to generalize. The goal in creating a linear regression model, therefore, is not to achieve perfect accuracy, which is impossible, but to make the model make predictions that are accurate to a reasonable degree.

II. METHODOLOGY

The data set used for the study is comprised of 4,747 entries contained in a .csv file, each containing 11 features. These features include date of posting, bedroom, hall, and kitchen (BHK) count, rent cost, size in square feet, floor location, area type, locality, city, furnishing status, preferred tenants, bathroom count, 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 2.1 Initial data set

Feature engineering was performed as a part of the data preprocessing procedure. Out of 12 features in the data set, only one (Posted On) was excluded, with the remaining 11 features being used for the model. Some of the features like size and BHK count could be used by the model without the need for any manipulation because they were already in numeric format, but in other cases, performing additional manipulations was necessary.

For instance, one-hot encoding was performed on certain features to represent string values in a way that can be provided to the machine learning model. This process results in each string value within a column being transformed into a column of its own, which is filled with Boolean values, where the only entries are either True or False.

In addition to feature engineering, outlier detection and treatment was performed. This was done through the interquartile range (IQR) approach, where data points above and below the IQR (multiplied by a given threshold) were omitted from the data set.

Afterwards, the data set was split into training and testing sets, with 20% of the data set being used for training and 80% for testing. Regularization was performed to allow the model to generalize better and prevent it from overfitting the data. This was done with ridge regression, where the model's weights were squared and multiplied by a given penalty term.

III. EXPERIMENTS

3.1 Outlier Detection and Treatment

Outlier detection was done through visual examination with a distribution plot of rent generated using the Seaborn library. The generated distribution plot was observed to be highly skewed, with a long tail extending to the right, indicating the presence of outliers on the higher end of the distribution. The sparseness of data in the skewed tail of the distribution plot indicates that the number of rental listings in the data set with such high rental prices were few compared to most of the listings, whose rent prices followed a more normal curve.

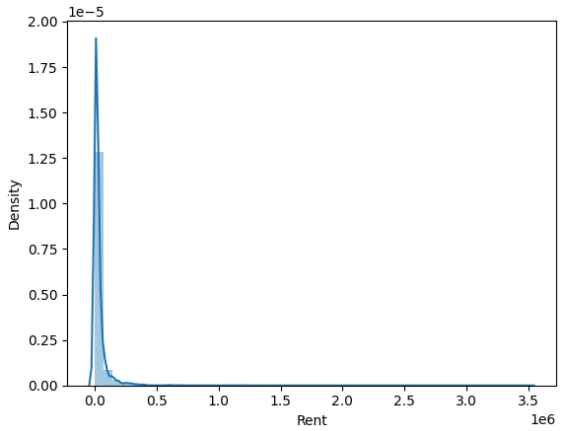

Figure 3.1.1 Distribution plot of rent prices before outlier treatment

Initial trials with the untreated data set showed suboptimal results, with the model's rent predictions on unseen data having poor R^2 (>0.5) and MSE scores. This was likely due to the more extreme rent prices skewing the model's predictions. As a result, it was necessary to remove data points with outlying rent prices from the data set.

To treat the data set for outliers, the interquartile approach was utilized. A standard threshold of 1.5 was chosen as the multiplier for the IQR to get the outlier range. Data points that were below the first quartile minus 1.5 times IQR were removed from the data set, as were data points above the third quartile plus 1.5 times the IQR.

The optimized data set with the outliers removed had a less significant skew and a more normal curve. Treatment with the IQR approach resulted in the removal of 520 data points, leaving the data set with 4226 data points.

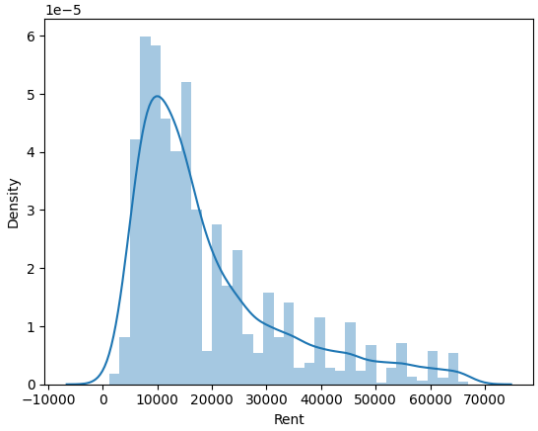

Figure 3.1.2 Distribution plot of rent prices after outlier treatment

With the data set treated for outliers, the model's rent prediction performance improved significantly, with its R^2 scores going beyond 0.5 and its MSE scores declining substantially in subsequent trials.

3.2 Feature Engineering

In order to improve the model's performance, features were selected and omitted from the data set. New features were also generated using existing features. Trials were conducted to determine how specific features affected the model's ability to make accurate predictions on rent prices.

3.2.1 Selecting Features

Out of the data set's original 12 features, only the Date Posted feature was not used. Choosing which features to include in the model proved to be, for the most part, a matter of trial and error. The exclusion of the Data Posted feature can only be justified by the difficulty of manipulating dates into a useful feature that could be understood by the model, and by intuiting that such a feature is likely inconsequential.

In the end, it was found that the remaining 11 features all proved helpful in optimizing the model's performance. The inclusion of even seemingly minute details like who to contact and preferred tenants improved the model's accuracy in rent price prediction, albeit to varying degrees.

The ability of other features in the model to increase its performance was more intuitive to comprehend. Quantitative information like size in square feet, floor count, and BHK count were good fits for the model since numerical data can be more easily used to map out a relationship with the target variable, i.e., rent, which is also a quantitative variable.

This is, of course, not to understate the usefulness of qualitative variables. For instance, variables like city and locality are particularly useful, as a rental's location is something that should be taken into account by the model because of the variability between rent prices in different locations. However, because such features were in string format, it was necessary to apply one-hot encoding in order to represent them in a way that could be interpreted by the model.

3.2.2 Splitting the Floor Feature

As the entries in the Floor column of the data set were mostly in string format, it was necessary to find a way to represent them in a format more amenable to the requirements of the model. This was done by splitting the column into two distinct columns: Floor Location and Total Floors. For most entries in the Floor column, this task proved trivial as most of them followed a format of "X out of Y."

```
Floor
X out of Y                  4708
basement level out of Y       34
unclear                        4
Name: count, dtype: int64
```

Figure 3.2.1.1 Counts of different representation formats for the Floor feature

Representing basement-level rentals numerically, however, proved to be a predicament because of the ambiguity of whether they were included in the total floor count. In the end, it was assumed that basement levels were not included in the given floor count. Working with this assumption, upper basement rentals were represented as 2 and lower basement rentals as 1, with the total floor count incremented by 2.

Another source of ambiguity showed in edge cases where the given floor location of the rental was a single digit, not following the usual format of "X out of Y". This problem was addressed by working with the assumption that the given floor location was the top floor.

Floor Location	Total Floors
1	2
1	3
1	3
1	2
1	2
...	...
3	5
1	4
3	5
23	34
4	5

Figure 3.2.1.2 A preview of the data set with the rental's floor location and the building's total floors separated

The creation of these two features benefited the model's performance, with an increase of 0.02 for the model's R^2 score and a drop of 3,606,478.78 for its MSE score.

3.3 Regularization

To address overfitting issues in the model, where the model has trouble generalizing and fits too closely to the training data set, regularization was performed. This was done through the use of ridge regression, where the model's weights are squared and multiplied by a penalty term α. This approach's performance, however, depends greatly on using the appropriate penalty term, so it is necessary to find the appropriate value to achieve optimal results.

For this, cross-validation was performed to find the appropriate penalty term. Through this process, it was determined that the appropriate penalty term was $\alpha = 63.0957344480193$. Using a penalty term of $\alpha = 100$ as a baseline, cross-validation improved the model's performance in terms of its R^2 score by 0.02 points.

IV. RESULTS AND DISCUSSION

4.1 Final MSE and R^2 Scores

Mean squared error (MSE)	Coefficient of determination (R^2)
46358723.18	0.76

Figure 4.1.1 MSE and R^2 scores of the model after

The final version of the model, with all the previously detailed procedures, resulted in an MSE score of 46,358,723.18 and an R^2 score of 0.76. While these scores clearly show that the model is not perfect, the

model's R^2 score is a satisfactory score that fits the study's purposes well.

4.2 Scatterplot Analysis

Figure 4.2.1 A scatterplot of actual rent prices and rent prices predicted by the model

As observed in Figure 4.2.1, the model's predictions are less distant from the true values on the lower end of the graph. This indicates that the model can predict the rent prices of lower-valued rental listings more accurately than it can predict those of higher-end rental listings. It is possible that this is because there are more data points where rent prices are on the lower end, leading to the model having trouble with higher-end listings where there is relatively less data. However, while there is less of it on the lower-end rental listings, it is clear that there is still quite a bit of divergence throughout the testing data set between actual and predicted rent prices.

It can also be observed from Figure 4.2.1 that the model, at least with rentals with prices above the median, has a greater tendency to underestimate rent prices rather than overestimate them. Additionally, while distant estimates can be observed throughout the graph, there is a more marked difference between predicted and actual rent prices among rentals with prices above the median.

V. CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary of Findings

With a final R^2 score of 0.76, the model, while imperfect, can predict the rent price of a rental listing with a reasonable degree of accuracy. While not quite close to 1, an R^2 score that indicates perfect accuracy, an R^2 score of 0.76 is considered substantial (Henseler et al., 2009).

However, an MSE score of 46,358,723.18 clearly leaves much to be desired. This indicates that there is much room for improvement for the model's ability to predict rent prices. Considering the tendency of the model to underestimate rent prices, especially with higher-end properties, this could prove problematic and cause a business utilizing the model to rent out properties with rent prices well below recommended rates. Conversely, since

In experimenting with which features to include and exclude from the model, it was found that all of them seemed to improve the model's accuracy in some way, albeit to varying degrees. Of these, quantitative features like BHK count and size in square feet seemed to have the greatest impact, possibly because it is less challenging for the model to deal with numerical values. Locational features like city and locality also had a great deal of impact, perhaps due to the fact that such features can serve as indicators for things like crime rate, pollution, availability and quality of public services, etc.

Other factors like furnishing status and area type also had a substantial positive effect on the model's performance, as such features are tied directly with the quality, and thus, the rent price, of a property being rented out. Outside of these, however, other features that were included in the model, like point of contact, had positive effects on the model's performance for reasons that are not obvious and could not be explained or justified easily. The positive impact of the inclusion of such features in the model clearly warrants further investigation into their relationship with rent prices.

5.2 Recommendations

Based on the findings of the study, it can be said that the linear regression model can predict rent prices with a reasonable degree of accuracy. This is especially true for lower-end rentals, where the difference between actual rent prices and the model's predictions are less dramatic and more consistent. However, it is still an imperfect model that can make overestimations and underestimations of rent prices, with varying degrees of severity. This assessment is particularly true for higher-end rentals above the median.

As such, while it can be used to predict the prices of lower-end rentals reasonably accurately, it is not as reliable for pricing higher-end rentals, e.g., rentals in more affluent areas, more spacious rentals, etc. Perhaps this could be addressed by training the model on a data set where the "desirability" of the rental properties is more evenly distributed, compared to the

data set used for this study, which seems to have a higher concentration of less desirable, lower-end rentals.

VI. References

Boston University. (n.d.). *InterQuartile Range (IQR).* https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_summarizingdata/bs704_summarizingdata7.html

Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. *Advances in International Marketing*, *20*, 277–319. https://doi.org/10.1108/s1474-7979(2009)0000020014

Kumar, A. (2023, February 14). *Linear Regression Explained with Real Life Example*. Data Analytics. https://vitalflux.com/linear-regression-real-life-example/#Real-world_examples_of_linear_regression_models

Machine Learning Department, Carnegie Mellon University. (2020, August 31). *The Overfitting Iceberg*. Machine Learning Blog | Carnegie Mellon University. https://blog.ml.cmu.edu/2020/08/31/4-overfitting/