Rent Price Prediction Model For India Using Linear Regression

Machine Learning CS-ELEC2C First Laboratory Activity

Lorenzo, Lex Zedrick M.

I. INTRODUCTION

One of the most common problems in real life is predicting a variable's value (the dependent variable) based on other factors of interest (the independent variables). This problem is solved by applying Linear Regression analysis. This method aims to minimize the discrepancies between the predicted and actual outcomes by fitting the variables in a straight line.

It is essential to remember that some facts in our case—the challenge of using linear regression to predict rent prices based on other provided data—must be more quantitative. As a result, we need first to transform the mentioned data into a quantitative format. Furthermore, there are disadvantages to linear regression despite its widespread usage and robustness. This is mainly observed when it is applied to resolve practical problems. Data collected in real-world situations occasionally deviates from a straight line, introducing complexity and variables outside the fundamental assumptions of the model.

II. METHODOLOGY

The dataset used in the study consists of 4,747 rows that are kept in a CSV file, each containing 12 characteristics. The posting date, number of bedrooms, halls, and kitchens (BHK), rental price, size in square feet, floor location, area type, area of locality, city, furnished status, preferred tenants, number of bathrooms, and point of contact are all included in these aspects.

Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
2022-06-06	1	200000	8000	Ground out of 4	Super Area	Beeramguda, Ra	Hyderabad	Unfurnished	Bachelors/Family	10	Contact Owner
2022-07-06	4	400000	7000	Lower Basement	Carpet Area	Jubilee Hills	Hyderabad	Semi-Furnished	Bachelors/Family	6	Contact Agent
2022-05-24	3	100000	6000	9 out of 12	Super Area	Navalur	Chennai	Semi-Furnished	Bachelors/Family	4	Contact Agent
2022-05-31	4	250000	5700	24 out of 25	Carpet Area	whitefield	Bangalore	Semi-Furnished	Bachelors/Family	4	Contact Agent
2022-06-01	4	1200000	5000	4 out of 15	Carpet Area	Juhu	Mumbai	Semi-Furnished	Bachelors/Family	4	Contact Agent
2022-07-02	4	130000	4800	11 out of 19	Carpet Area	whitefield	Bangalore	Semi-Furnished	Bachelors/Family	4	Contact Agent
2022-07-10	4	150000	4761	14 out of 18	Super Area	Kukatpally, NH 9	Hyderabad	Semi-Furnished	Family	5	Contact Agent
2022-06-11	4	200000	4761	1 out of 35	Carpet Area	K P H B Phase 4	Hyderabad	Semi-Furnished	Bachelors	6	Contact Owner
2022-06-18	4	150000	4530	3 out of 5	Super Area	Jakkur Plantation	Bangalore	Semi-Furnished	Family	5	Contact Agent
2022-05-27	5	600000	4500	7 out of 20	Carpet Area	Bandra West	Mumbai	Furnished	Bachelors	5	Contact Agent
2022-06-30	4	160000	4500	4 out of 4	Carpet Area	Langford Garden	Bangalore	Unfurnished	Bachelors	4	Contact Agent
2022-06-28	6	280000	4500	Ground out of 1	Carpet Area	Raja Annamalai I	Chennai	Semi-Furnished	Bachelors	5	Contact Agent
2022-06-30	4	250000	4500	Ground out of 2	Carpet Area	Jubilee Hills	Hyderabad	Semi-Furnished	Family	4	Contact Agent
2022-07-06	5	80000	4500	1 out of 2	Super Area	Kukatpeya Nagar	Hyderabad	Semi-Furnished	Bachelors/Family	6	Contact Owner
2022-07-06	4	130000	4341	9 out of 15	Carpet Area	Arakere	Bangalore	Semi-Furnished	Bachelors/Family	4	Contact Agent
2022-07-06	6	30000	4200	Ground out of 2	Super Area	Kuntloor	Hyderabad	Semi-Furnished	Bachelors/Family	5	Contact Owner
2022-06-08	4	120000	4165	11 out of 25	Carpet Area	Hebbal	Bangalore	Semi-Furnished	Bachelors/Family	4	Contact Agent
2022-07-02	5	300000	4050	Ground out of 5	Super Area	Bandlaguda Jagir	Hyderabad	Unfurnished	Bachelors/Family	4	Contact Owner
2022-07-04	4	30000	4000	Ground out of 3	Super Area	Salkher Bazar, Pul	Kolkata	Furnished	Bachelors/Family	3	Contact Agent
2022-06-08	4	190000	4000	Ground out of 2	Super Area	Koramangala	Bangalore	Furnished	Bachelors/Family	5	Contact Owner
2022-06-11	4	530000	4000	3 out of 4	Carpet Area	Anand Niketan	Delhi	Semi-Furnished	Bachelors/Family	5	Contact Agent
2022-06-28	4	150000	4000	3 out of 3	Carpet Area	West End	Delhi	Semi-Furnished	Bachelors	7	Contact Agent
2022-07-08	4	120000	4000	9 out of 9	Super Area	Perungudi	Chennai	Semi-Furnished	Bachelors	4	Contact Agent
2022-07-10	4	220000	4000	1 out of 2	Super Area	Poes Garden, Tey	Chennai	Semi-Furnished	Bachelors/Family	4	Contact Agent
2022-07-09	5	500000	3900	4 out of 6	Carpet Area	Altamount Road	Mumbai	Furnished	Bachelors/Family	5	Contact Agent

Figure 2.1 Data Set used for the model

As a part of feature engineering and preprocessing of the data, It has been decided that 10 of the features will be used for the model. The "Posted on," "Area of Locality," "Bathroom," and "Area Type" have been excluded because, during the experiment, they either lessened the model score or were deemed insignificant. During data exploration, it can be seen that some features had inconsistent rows, while some features, like the City and Furnishing status, can not be directly used in the model. Therefore, some data manipulation is needed to be used in the model.

Data cleaning was primarily done in the "Floor" column of the dataset. Data inconsistency was addressed by correcting the naming convention of the ground and basement floors, switching the house floor and the total number of floors, and adapting the format to suit the model.

One hot encoding addresses the features that need to be quantified. This method creates separate columns for the unique instances of the feature that need to be manipulated for it to be inputted in the data model. The separate columns will then be filled with boolean values to identify each row uniquely.

Outlier detection and removal were also done on the target (the rent). Removal of outliers is necessary because it prevents the outliers from affecting the estimation of the model. The detection and removal were implemented using z-scores to calculate which data points deviate too far from the mean.

III. EXPERIMENTS

3.1 Feature Engineering

Feature engineering is transforming raw data from datasets into features suitable for machine learning models—in other words, adding, deleting, creating, and transforming the most suitable features from the data at hand to make the machine learning models more accurate and efficient.

3.1.1 Selecting features

This machine learning model's features were selected mainly by trial and error. Adding and deleting features that significantly change the R^2 and Mean Squared Error.

The decision to exclude the "Posted On" column is mainly because of the difficulty and insignificance of the date the house is posted on to the prediction of house prices. The "Bathroom," "Area Type," and "Area Locality" are excluded because these features did not have that many significant changes in the overall test score of the model. It just made the model have to calculate complex, unnecessary relationships. The remaining eight features proved to help maintain a balance of high training and test scores.

3.2 Data cleaning for the selected features

3.2.1 Data cleaning for floor column

In the data cleaning process, the "Floor" column is the one that went with the most refinement. During data exploration, inconsistencies in the naming convention for the ground and basement floors were the easiest to spot, and things like there is no total floor in some data points and some points where the house floor is higher than the total floor were the ones that needed proper searching.

The representation of the data points where a number does not represent the house floor needs to be equated with a proper number equivalent to the house floor. The first thought was to equate it with negative values, which resulted in the evaluation being biased against the basement floors. Then, I decided to count it as part of the total floors and gave the upper basement a value of 2 and the ground and lower basement a value of 1. This lessened the bias against the basement floor while removing the negative values in the evaluation.

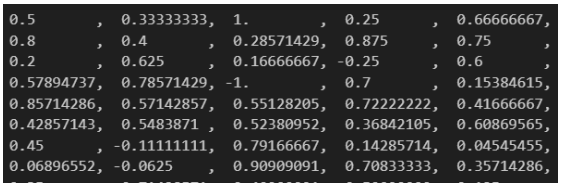

Figure 3.1 Negative evaluation of "Floor" values

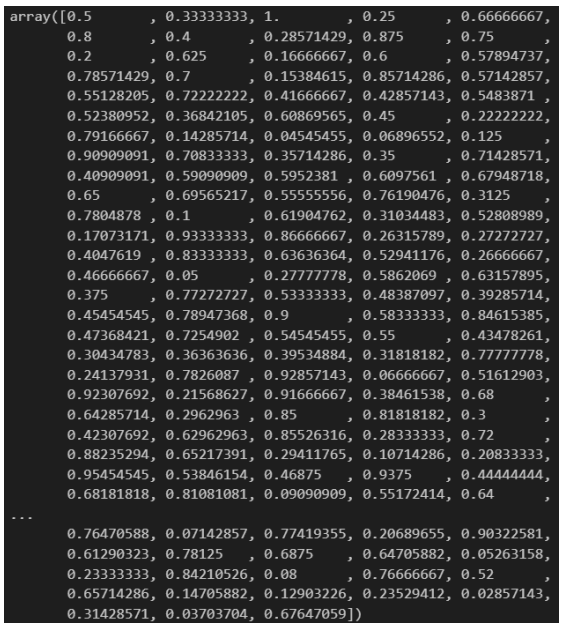

Figure 3.2 Solved the issue of negative values for "Floor"

The next problem I addressed was that some data points had no total floors; it was just a number or "Ground." In this case, I assumed that this instance means that the house floor is the top floor. Therefore, I replaced these data points to be equivalent to 1.

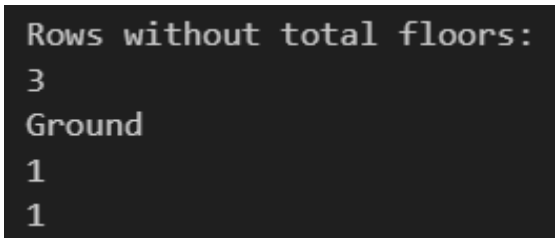

Figure 3.3 List of data points without total floors

The last problem addressed for the "Floor" column is that some data points on the house floor had higher values than the total floor. I assumed this was caused by human error; therefore, I switched the places of the two numbers.

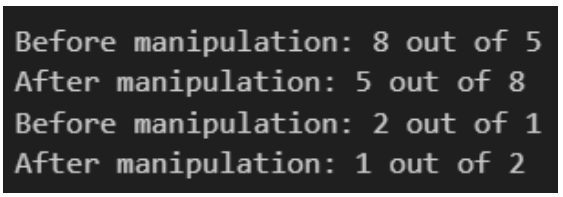

Figure 3.4 Switched the places of house floor and total floor

Now that the data of the "Floor" column had been cleaned, I evaluated each data point by replacing " out of " with "/" and then evaluated the string using a function.

3.2.3 Removing Outliers

Outliers are typically removed for machine learning models because these values significantly affect the model's prediction. In this prediction model, the rent prices are scored using a z-score, which measures the standard deviation of each date from the mean. The threshold set for the measurement is 1.7. This allows for some leniency while also reducing the number of outliers.

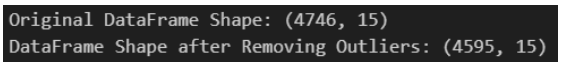

Figure 3.5 Results of removing outliers

3.3 Splitting of data

The splitting proportion for this model is 80% for the training set and 20% for the test set. This data division allows the model to learn more about the relationships between the features while maintaining a significant number for the test set.

3.4 Regularization

Regularization is done so that the model can generalize better and prevent it from overfitting. This model uses a data Library called Ridge Cross-Value to find the best alpha/penalty value multiplied by the

model's weight squared. The optimal alpha will then be inputted into a Regularization model Library called Ridge.

IV. RESULTS AND ANALYSIS

4.1 Final MSE, training, and R^2 scores

Mean Squared Error	153,225,096.37
Training Score	0.75
R^2 score	0.80

Figure 4.1 Final results of the model

After all the procedures, the model results are 153,225,096.37 for the Mean Squared Error (MSE), 0.75 for the training set, and 0.80 for the test set. Although the MSE is relatively high, the training and test set scores are considerably high but could be better. The closeness of the training and test scores means there is no overfitting, and the model has done quite well in predicting the prices for those sets that it had not seen.

4.2 Scatter Plot Analysis

Figure 4.2

Figure 4.2 shows that the predictions on the lower left quadrant are closer to the actual rent prices than those of the upper right quadrant. This entails that the model can more accurately predict the prices of those low-value listings compared to high-value ones. This may be because of the nature of the given dataset.

V. CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusion

While the final R^2 score of 0.80 indicates that the model has a high accuracy in predicting rent prices, it is also essential to consider the MSE, which is significantly high. This means that the model still needs improvement, and improvements can still happen. The model's tendency to underestimate the prices of high-value houses can cause problems when predicting the prices of those particular houses.

During the experimentation stage, it was found that not all features proved to be helpful. For instance, the "Area Locality" and "Area Type" proved to have lessened the test scores. Moreover, seemingly insignificant features like "Point of Contact" had one of the most significant impacts on prices. This may be because agents have to increase the prices for their commission, while the owners do not have to do this because they will pocket the rent for themselves.

5.2 Recommendations

To further enhance this model's accuracy, adding factors that may affect rent prices, like the neighborhood status, whether the area around the house is safe, or if amenities are offered, is recommended. The house's proximity can also be considered for this model, whether it is close to a public

transportation network or if it is close to schools, office areas, and such.

VI. REFERENCES

The Magic of Linear Regression Model. (n.d.). Www.linkedin.com. Retrieved February 4, 2024, from https://www.linkedin.com/pulse/magic-linear-regression-model-bhagyashree-ghosh/

What is One-hot Encoding. (n.d.). Deepchecks. Retrieved February 4, 2024, from https://deepchecks.com/glossary/one-hot-encoding/#:~:text=One-hot%20encoding%20in%20machine

What is Feature Engineering? (2023, March 20). GeeksforGeeks. https://www.geeksforgeeks.org/what-is-feature-engineering/

Barkved, K. (2022, March 9). How To Know if Your Machine Learning Model Has Good Performance | Obviously AI. Www.obviously.ai. https://www.obviously.ai/post/machine-learning-model-performance