Analysis and Prediction of Real Estate Rent

Project Report for Indian Institute of Technology, Bombay - DS203: Programming for Data Science (2022)

Anway Deshpande
Dept. of Electrical Engineering
Indian Institute of Technology, Bombay
Mumbai, Maharashtra
210070011@iitb.ac.in

Anuj Gupta
Dept. of Electrical Engineering
Indian Institute of Technology, Bombay
Mumbai, Maharashtra
21D070014@iitb.ac.in

Samar Agarwal
Dept. of Electrical Engineering
Indian Institute of Technology, Bombay
Mumbai, Maharashtra
21D070059@iitb.ac.in

Abstract—**This document is a report of our project aiming at predicting Real Estate Rent. We all face various issues when we change places for better opportunities or when we begin the journey of Entrepreneurship, one being finding a suitable place on rent. Our project aims at studying various parameters which directly or indirectly impact real estate rent. Moreover, we have demonstrated the dataset using various plots and graphs which provides the necessary intuition of all the factors governing house rents in metropolitan cities.**

I. Introduction

More than 50 percent of the world's population lives in urban areas and it is predicted that the number will rise to 75 percent by the 2050s. With constant urbanization and population growth, the demand for housing keeps on increasing. But resources, mainly land being limited, demand for buying houses will reduce and people will start renting and living in rented houses.

So, for new migrants in the city, the most excruciating part is to find good houses with affordable rents. And without proper knowledge of different factors governing the rent prices, they end up taking houses that do not justify their rent. Therefore, this is a major problem that may hamper individual growth and he/she might not live up to his/her expectations. Therefore it is necessary to get an insight into the rents at different localities in various cities. This is exactly what our model aims to do.

In this project we have primarily done the following things:

- We have collected data from various sources including Kaggle which is a subsidiary of Google LLC.
- We have performed extensive data cleaning and filtering to reduce a large dataset to a compact and easily usable form.
- We have analysed the impact of various factors including locality, number of bedrooms and bathrooms, type of owners and property type being furnished, semi-furnished or non-furnished on the rent.
- Starting with basic Exploratory Data Analysis on our data set, we made various plots and graphs to practically manifest our data set for better intuitions.
- We have used basic machine learning techniques including linear, ridge and lasso regression for prediction.

II. Dataset and Preproccessing

We acquired the data from various sources including Kaggle and then preprocessed the data to make it clean. Our data includes housing data of various well-known cities on which we performed basic EDA and then ran basic ML framework for actual prediction.

A. Datasets

Our data includes housing data of 8 different well to do cities which are Mumbai, Delhi, Ahmedabad, Bangalore, Chennai, Delhi, Hyderabad, Kolkata and Pune. We are aimed to predict the real estate rent considering various parameters including area, furnishing status, number of bedrooms and bathrooms. Moreover, we have data containing crucial impact of types of seller (”agent” or ”seller” or ”owner”), types of layout(”BHK” or ”RK”), types of property(”Apartment”, ”Studio Apartment”, ”Independent House”) and locality of that estate(different set for different city). The only shortcoming is the limited number of cities we have data on which might make our model inefficient for predicting rent of a general city.

B. Preprocessing

We have roughly ten thousands of data for each city which sums up to 2 lakhs total data. Though, we have data of each city separately but for the purpose of analysing and studying them we merged the datas into a single dataset. Moreover, we added a new column containing city names to account for the source of each data.

Our preprocessing included the following steps-

- Merging eight different city dataset into one single dataset.
- Checking data types and null values in each column.
- Dropping comma contained in the prices so as to convert it into float datatype to perform EDA.
- Converted ”**bathroom**” and ”**price**” column from object type to ”**float64**”.
- Multiplying the price not containing comma with 100000 make it reasonable.
- Removing outliers from our dataset.

III. Data Visualisation

For best intuitions, we made several plots including different parameters to get a deep understanding of the interdependence of rent on those factors.

A. Box and Whisker Plot

*1) **log(prices) vs city**:* We created a separate column containing log of prices to get a better and stretched box and whisker plot. The plot between log of prices and city can be found in figure 1.

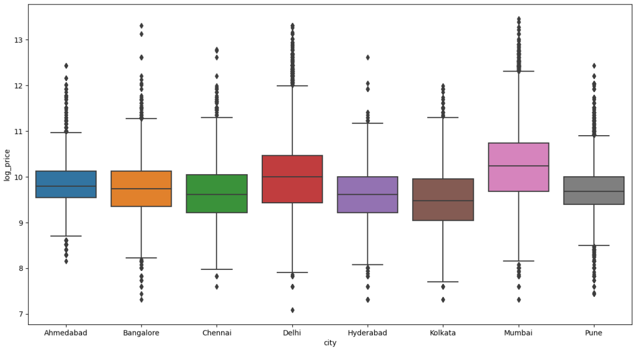

Fig. 1. log(prices) vs city

Observations

- Median price of all cities is almost same. Though, Mumbai wins if we observe accurately.
- Mumbai and Delhi are among the top when considering maximum price in the cities according to the data.
- With no surprise, Delhi is also among the cities with lowest rent. Kolkata being the lowest.

*2) **log(prices) vs seller-type**:* The plot between log of prices and seller-type can be found in figure 2.

Observations

- Median of rent follows the trend of Agent charging the highest rent followed by Builder and then Owner.
- Considering the highest rent charged the trend follows Agent, then Owner and then Builder.
- Surprisingly, observing the minimum rent charged we found that Owner charged least, then Agent and then Builder.

*3) **log(prices) vs furnish-type**:* The plot between log of prices and furnish-type can be found in figure 3.

Observations

- Analysing maximum, median and minimum rent, we found that all are following same trend. Furnished having the highest rent, then semi-furnished and unfurnished having the lowest rent.

*4) **log(prices) vs no. of bedrooms**:* The plot between log of prices and number of bedrooms can be found in figure 4.

Observations

- On analysing, we found that as number of bedrooms increases, the rent price also increases.

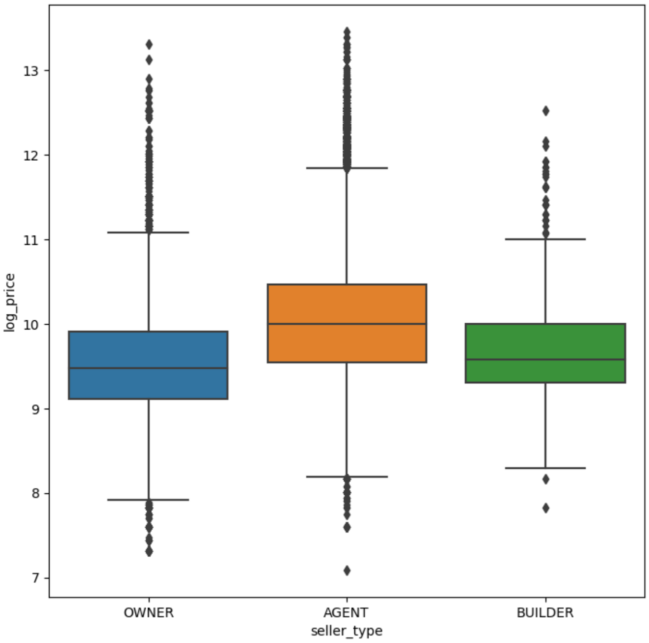

Fig. 2. log(prices) vs seller-type

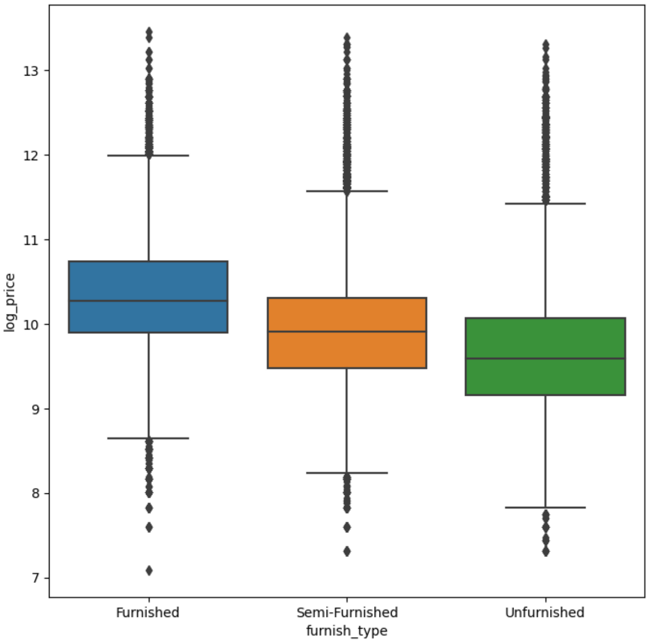

Fig. 3. log(prices) vs furnish-type

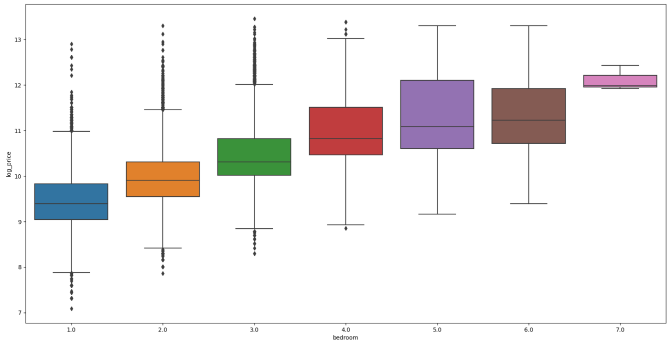

Fig. 4. log(prices) vs number of bedrooms

B. Scatter Plot

We created a separate column containing log of prices to get a better and non congested scatter plot. The plot between log of prices and city can be found in figure 5.

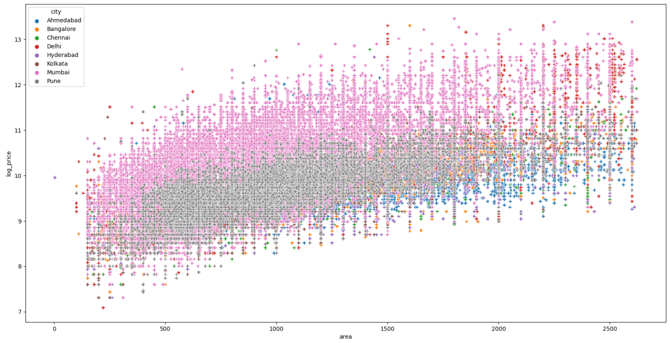

Fig. 5. Scatter Plot

Observations

- As always Mumbai dominates the rent and is the one charging the highest price.
- Delhi holds a significant share in large area houses, sometimes even surpassing Mumbai.
- Pune follows Mumbai in mid-price range, followed by Chennai and Hyderabad.
- Ahmedabad started showing its dominance in large area houses, though after Mumbai and Delhi.

C. Correlation Matrix

We plotted a heatmap to show inter-dependence of various parameters governing rent. The heatmap or the correlation matrix can be found in figure 6.

Observations

- As expected, price and area are positively correlated and they have almost equal correlation of around 0.45.
- Number of bathrooms and bedrooms are also positively correlated with price with almost same correlation.
- With no surprise, as area increases, number of bedrooms and bathrooms also increases as they are highly correlated with correlation of 0.85.

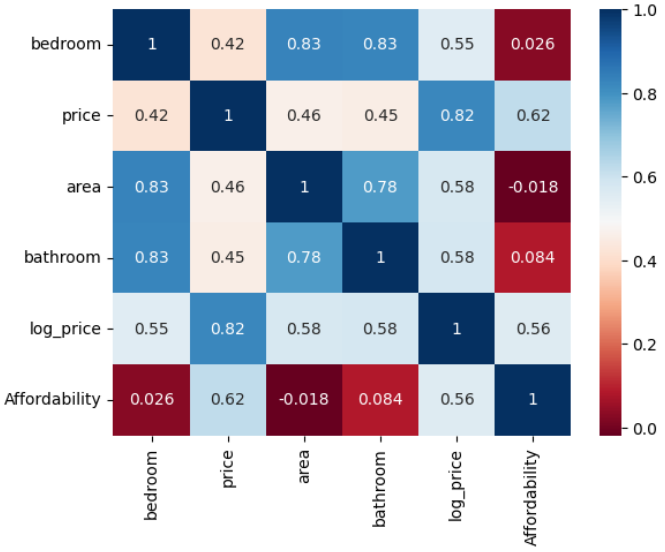

Fig. 6. Heatmap

D. More interesting graphs

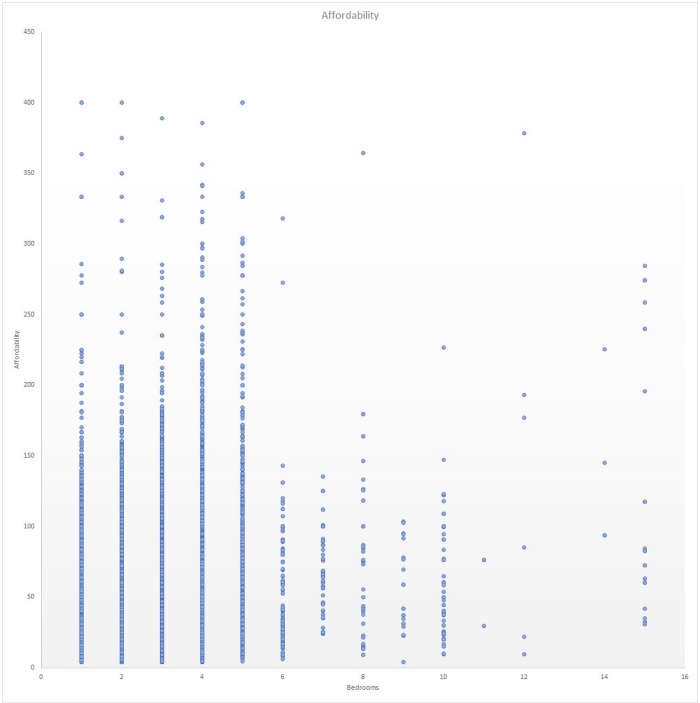

Fig. 7. Scatter Plot between area and affordability

IV. Machine Learning Modelling

We have used the most basic machine learning model i.e. Regression. We first splitted the data into train and test and then applied three regression models, they are as follows :-

- Linear Regression (OLS)
- Ridge Regression
- Lasso Regression

A. *Linear Regression*

The Residual Plot of Train Samples and plot between y_test vs y_pred_test for this regression is given in figure 8.

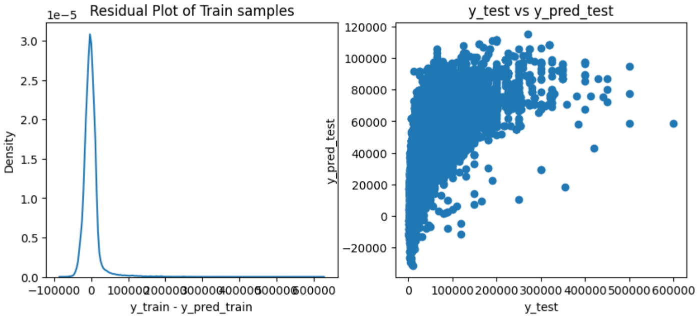

Fig. 8. Linear Regression

Observations

- y_test vs y_pred_test is deviated from an ideal straight line, thus the model is inaccurate.

B. *Ridge Regression*

The Residual Plot of Train Samples and plot between y_test vs y_pred_test for this regression is given in figure 9.

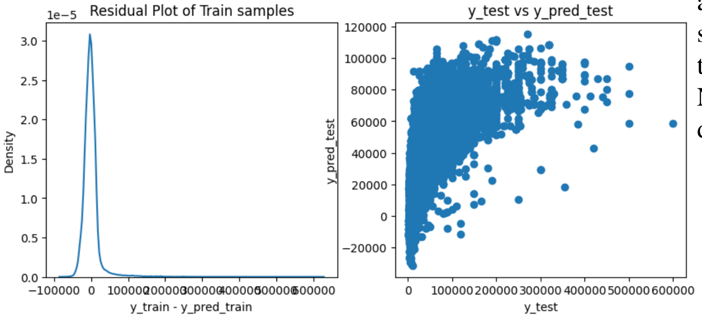

Fig. 9. Ridge Regression

Observations

- y_test vs y_pred_test is deviated from an ideal straight line, thus the model is inaccurate.

C. *Lasso Regression*

The Residual Plot of Train Samples and plot between y_test vs y_pred_test for this regression is given in figure 10.

Observations

- y_test vs y_pred_test is deviated from an ideal straight line, thus the model is inaccurate.

The table containing the R2 Score of train as well as test data and Cross_val mean of train data is given in table I.

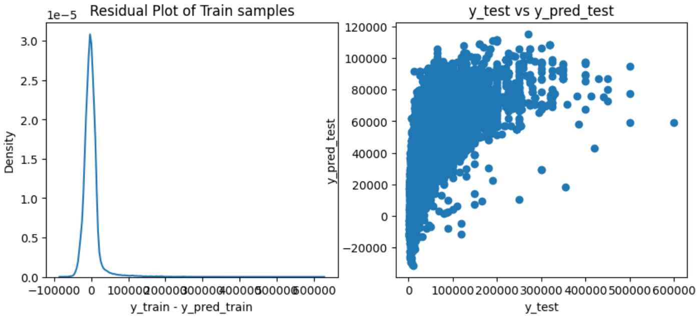

Fig. 10. Lasso Regression

TABLE I
R2 SCORE AND CROSS_VAL MEAN

Models	**R2 Score(Train)**	**R2 Score(Test)**	**Train CV Mean**
Linear Reg.	0.41	0.42	0.41
Ridge Reg.	0.41	0.42	0.41
Lasso Reg.	0.41	0.42	0.41

V. CONCLUSION

This project taught us a lot about data analysis and how it can be effective in solving real life problems. We kept the model extremely simple which inevitably triggered inaccuracy since we wanted to focus mode on Exploratory Data Analysis. In future, we could use higher machine learning framework and include various otherwise overlooked parameters like social beliefs of people, locality and much more to increase the accuracy of our model.
Nevertheless, The project experience for all of us was tremendous and we learnt a lot through the process.

REFERENCES

[1] https://www.kaggle.com/datasets/saisaathvik/house-rent-prices-of-metropolitan-cities-in-india

[2] https://www.kaggle.com/code/saisaathvik/rent-price-prediction-for-all-cities

[3] https://www.kaggle.com/code/rajpraveenpradhan/cp-prediction-with-model-comparision-deployment