Assignment 12.2

Project Milestone – 5: Final Project Paper.

Sucharitha Puppala

Data Science, Bellevue University

DSC630-T302 Predictive Analytics (2231-1)

Prof. Andrew Hua

November 12, 2022.

Abstract

In this data mining project, machine learning is used to predict the house rental price in India. This project helps the rental sites to update the details of the houses that are available for rent as per the preference of the people, based on the previous data collected for example number of bedrooms, hall and kitchen, number of bathrooms, size of the house etc. In this paper the various features that are considered for rental house in India are identified. CRISP – DM methodology is followed for the project. As the project involves supervised learning with continuous target variable, four machine learning models are selected and evaluated and discussed whether the model is ready for deployment or not.

Final Project

1. Introduction

Rental housing is an integral part of the housing tenure systems in cities, and is also integral to the stages of a migrant's upward mobility from squatter settlement to ownership housing. An examination of the residential rental housing situation in India during the last decades using data from the Census of India and the National Sample Surveys finds that more than one-tenth of the households in India lived in rented houses in 2011, of which almost four-fifths of the total households living in rented houses in India were in the urban sector. Moreover, while the issues of shelter deprivation of many households and the question of affordability of shelter remain, a new phenomenon of a sharp rise in the number of vacant houses during the last decade has added to the severity of the housing problem.

The India rental housing market is driven by the growing influx of migrants from non-metro cities to metro cities for occupational and educational purposes. This has drastically increased the demand for affordable rental spaces in the proximity of the working spaces or educational institutions. This has also led to the emergence of the concept of co-living.

Additionally, increasing prices of land, houses and flats especially in Tier 1 cities is further expected to propel the market growth through FY2027. Furthermore, sometimes those who have the resources and can afford houses or land do not get appropriate investment opportunities, hence prefer living in a rented property, thereby fueling the market growth.

The demand for residential rental houses in top seven cities in India has increased by 10-20 per cent in 2022 as compared to the pre-pandemic period in 2019. In January-March 2022 (Q1 2022), total rental housing demand (searches) in 13 Indian cities jumped by about 15.8% quarter on quarter (QoQ) and 6.7% year on year (YoY), while the cumulative rental housing supply (listings) increased 30.7% QoQ and 101.5% YoY across the cities mapped.

Final Project

The India rental housing market is segmented based on type, property type, size of unit, location, company, and region. The Human Rights Measurement Initiative finds that India is doing 60.9% of what should be possible at its level of income for the right to housing.

The following project focuses on the features that are more considered in the rental houses. The results obtained from the analysis can be considered for identifying the type of houses that are more preferred for rentals. In this project analysis is done on the rental houses available in different cities in India.

This project implements the standard data mining methodology. Information discovered at various

Stages of Cross Industry Standard Process for Data Mining (CRISP-DM) are a foundation for future recommendations regarding the predictive potential of the dataset. Software applied in this project is Jupyter Notebook open-source software, open standards, and services for interactive computing across multiple programming languages.

2. CRISP DM

CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is an industry-proven way to guide your data mining efforts.

- As a methodology, it includes descriptions of the typical phases of a project, the tasks involved with each phase, and an explanation of the relationships between these tasks.
- As a process model, CRISP-DM provides an overview of the data mining life cycle.

It has six sequential phases:

- Business Understanding - The Business Understanding phase focuses on understanding the objectives and requirements of the project.

- Data Understanding - Adding to the foundation of Business Understanding, data understanding drives the focus collection of data, verification of data quality and data exploration.
- Data Preparation - This phase, which is often referred to as “data munging”, prepares the final data set(s) for modeling. It has five tasks: Select data, clean data, Construct data, Integrate data, and Format data.
- Modeling-This phase has four tasks: Select Modeling Technique, Generate Test Design, Build Model and Assess Model.
- Evaluation- In this phase we Evaluate Results, Review Process and Determine Next Steps.
- Deployment – In this phase we Plan Deployment, Plan Monitoring and Maintenance, Produce Final Report and Review Project.

3. Business Understanding

“House rent prediction “for the rental houses available in different places of India. This project will focus on the factors that influence in prediction of rent around various locations of India. The topic raises the following questions:

1. Which locations are having more rental houses in India?
2. What dataset features are helpful in the prediction of house rent?
3. Which types of houses are more preferred for renting in India?

4. Data Understanding

4.1 Dataset

The data set used for this analysis is House Rent Prediction dataset from kaggle.com. The data set contains the data of the houses available for rent in India. In this Dataset, we have information of almost 4700+ Houses/Apartments/Flats Available for Rent with different parameters like BHK, Rent, Size, No. of

Floors, Area Type, Area Locality, City, Furnishing Status, Type of Tenant Preferred, No. of Bathrooms, and Point of Contact.

The following is the link to the data set

https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset

4.2 Data Exploration

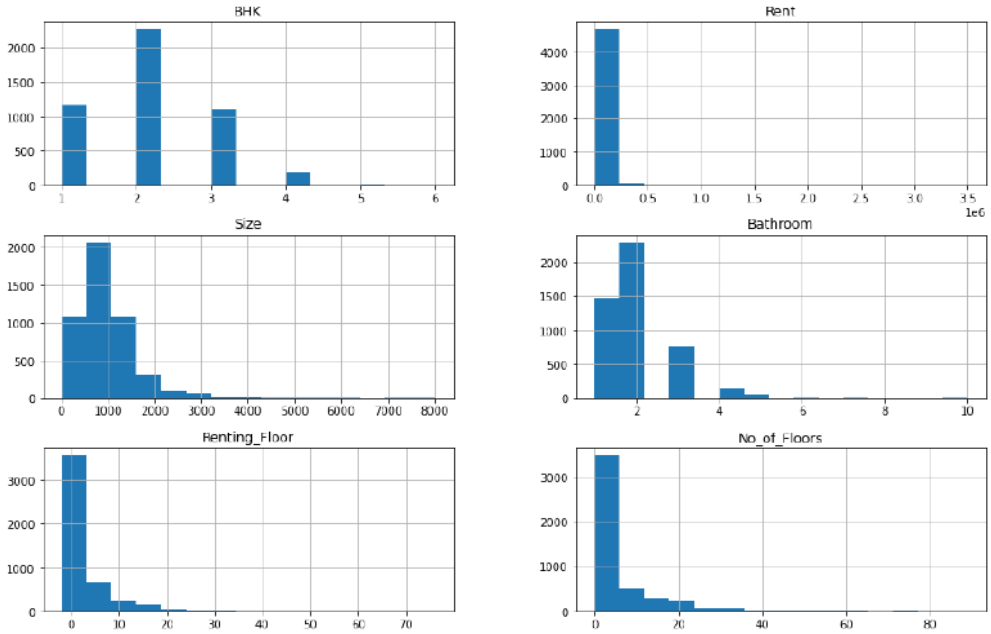
The dataset has 4746 rows and 12 columns. The dataset contains both numerical and categorical data.

There are no null values in the dataset. The number of unique values are extracted from the dataset variables, and observed many unique values in 'Area Locality' column, 'Size', 'Floor' variables. The target variable is "Rent" from the dataset. The 'Floor' column is split into two columns i.e. 'Renting_Floor' and 'No_of_Floors' for clear understanding and Ground, Upper Basement and Lower Basement values that are in string are replaced with numerical values, like Ground = 0, Upper Basement = -1 and Lower Basement = -2 for easy analysis.

For the exploratory data analysis Bar plots, Histograms, Boxplot, Pair plot and Correlation coefficient heat map are used.

a. Histograms are used for understanding the distribution of numerical variables of the dataset.

Figure 1: Histograms of the numerical features in the dataset.



From the Figure 1, we see that the distribution of ‘rent’ is skewed towards right, from the distribution of 'BHK'(Bedroom, Hall and Kitchen), we can see that 2 BHK houses are more in number among other numbers for renting, from the distribution of variable 'Size', we can see that 500 to 1100 square feet houses are more for renting, from the distribution of variable 'Bathroom', we can see that 2 bathrooms are more in number for renting, from the distribution of variable 'Renting Floor', we can see that mostly ground, first, second, third floors are more in number for renting, from the distribution of variable 'No_of_Floors', we can see that up to 10 floors are more in number for renting

b. Bar plots are used for understanding the count of the categorical variables distribution in the dataset. From Figure 2 we can say that the rents are more in Mumbai city when compared to other cities in India, Figure 3 indicates that as the number of BHK increases the rent increases.

Figure 2: Distribution of rent by city

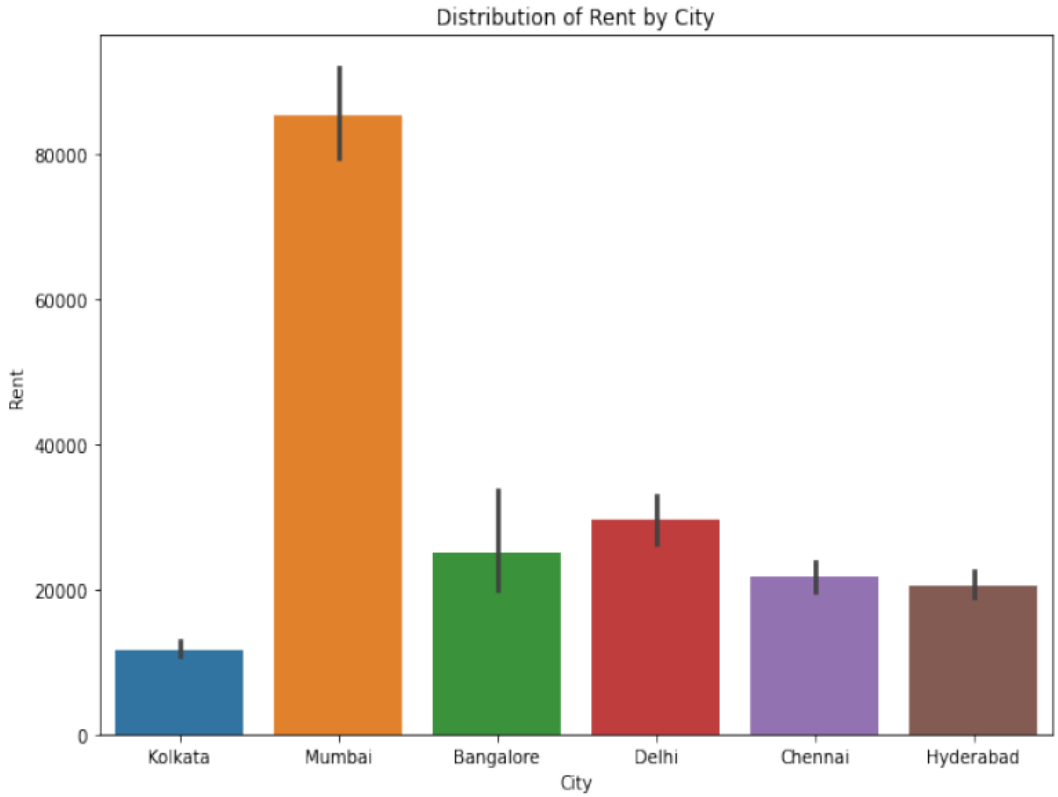
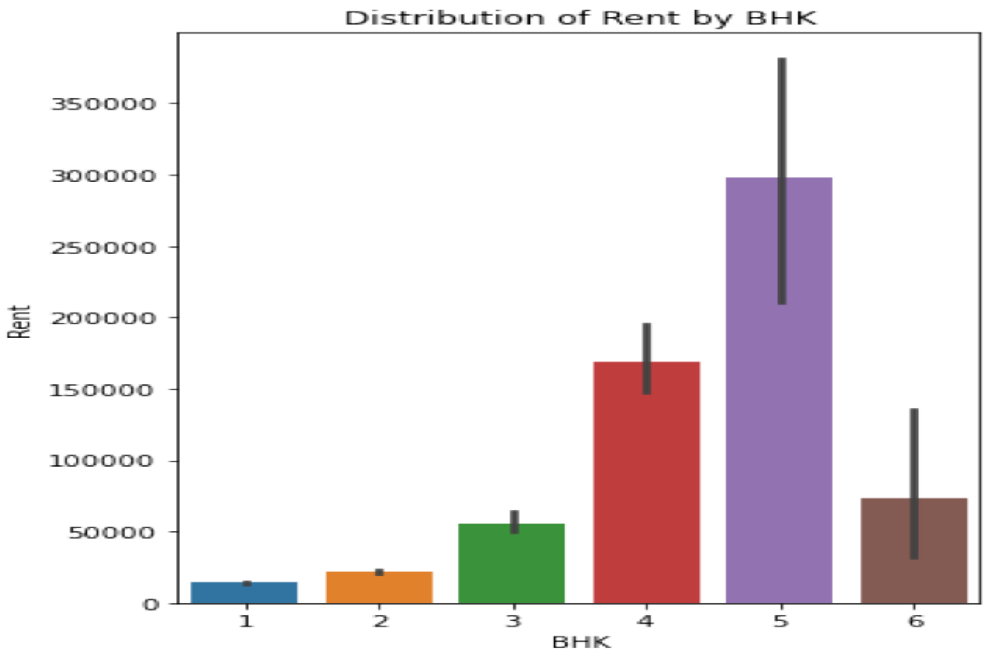
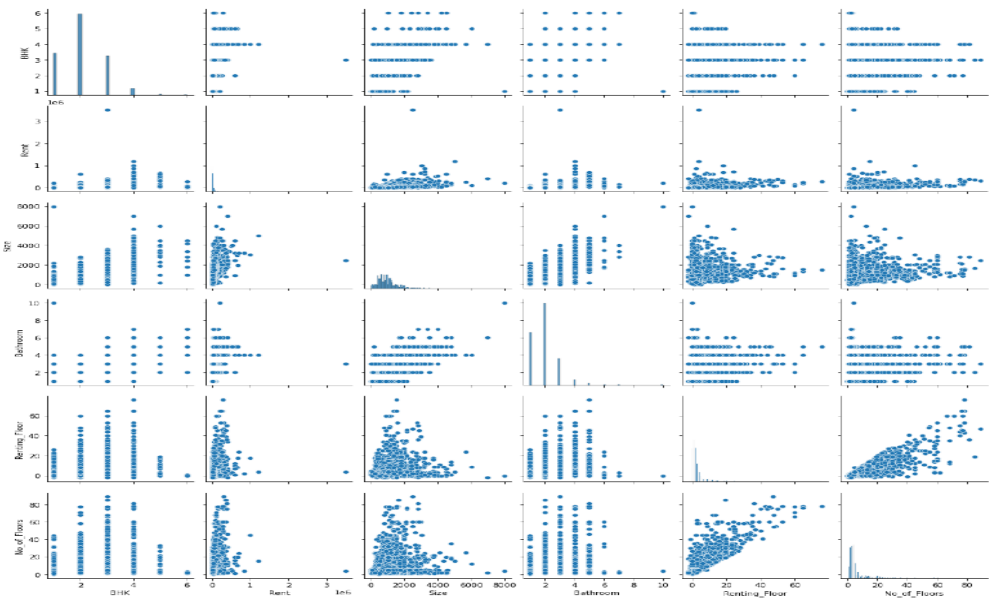


Figure 3: Distribution of rent by BHK



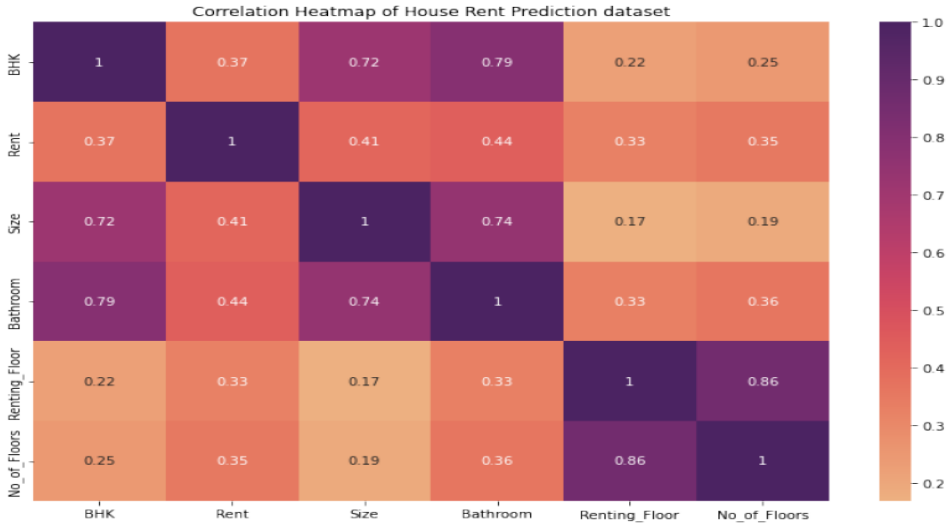
c. Pair plot is used to understand the best set of features to explain a relationship between two variables or to form the most separated clusters.

Figure 4: Pair Plot



d. Correlation Heat map is plotted for understanding the feature that is having good correlation with the target variable.

Figure 5: Correlation Heat map



e. Observations from Exploratory data analysis

When observed the count plots the rental houses with Super Area, Semi- Furnished house, Bachelors/ Family tenants preferred are available in more number. The count plots in Appendix-A will give clear idea about the distribution of the variables in the dataset. Point of contact preferred is Contact Owner for renting the houses. When observed the distribution of the rent with the different features in the dataset we see that rents are high for houses with “Carpet Area”, tenant preferred type as “Family”, point of Contact by “Contact Agent”. The bar plots from Appendix- B gives the distribution of the target variable ‘Rent’ with different categorical variables.

5. Data Preparation

The process of prepping the data involved the following steps.

1. Removed the unwanted columns from the dataset.

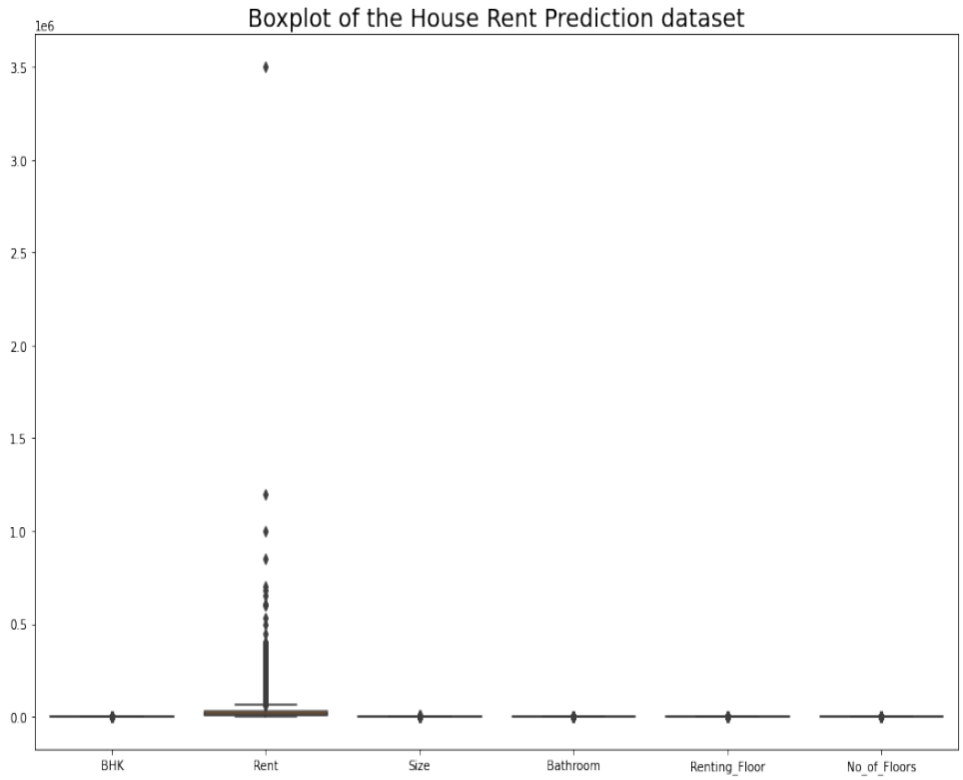
- Some of the columns from the dataset selected are not necessary during the analysis and are dropped from the dataset.
- The columns 'Posted on', 'Floor', 'Area Locality' are dropped from the dataset.
- ‘Posted on’ column is dropped as all the data collected is from 2022 and there is no impact on the analysis.
- ‘Floor’ column is dropped as we have split the column in two columns “Renting_Floor” and “No_of_Floors”.
- “Area Locality” column is dropped as there are many unique values and the project mainly focuses on the number of houses available for rent and the rental values in different cities in India but not a particular city.

- After dropping the unwanted columns, check for null values and duplicate values is done. If any null values or duplicated values are identified those values are dropped and the shape of the dataset is identified.

2. Identified the outliers present in the dataset using boxplots and are to be removed for the further analysis.

- In the analysis the rent column is found to have outliers, i.e. the rents which are greater than or equal to one lakh are identified and are dropped from the dataset as there are very few houses available for rental with one lakh and above. The Boxplots of the target variable 'Rent' before removal of outliers and after removal of outliers can be seen in Appendix- C for clear understanding.

  Figure 6: Box plot for outliers' identification



3. Creating Dummy Variables for the Categorical Features

- The categorical features present in the dataset are identified and dummy variables are created.
- This creation of dummy variables enables us to use a single regression equation for representing multiple groups.
- Check for null values is done and if any present are dropped from the dataset.
- After that the shape of the dataset is obtained.

4. Split the dataset into train and test datasets using 'Rent' as the target variable.

- Initially the dataset is split into features(X) with the target variable dropped from the dataset and target(y) variable by considering only the 'rent' column from the dataset.
- The data set is to be split into training and test datasets using 'Rent' as the target variable.
- Here the dataset is split into 80% training dataset and 20% test dataset, taking the test size as 0.2 and the random_state as 42.
- X_train, X_test, y_train, y_ test are created using train_test_split()
- The train and test data sets shapes are obtained for better understanding the split of the dataset.
- The features train and test datasets are standardized using StandardScaler.
- The X_train dataset is transformed and fit to the standard scaler and the X_test dataset is just transformed but not fit.
- With this the data is ready for model building.

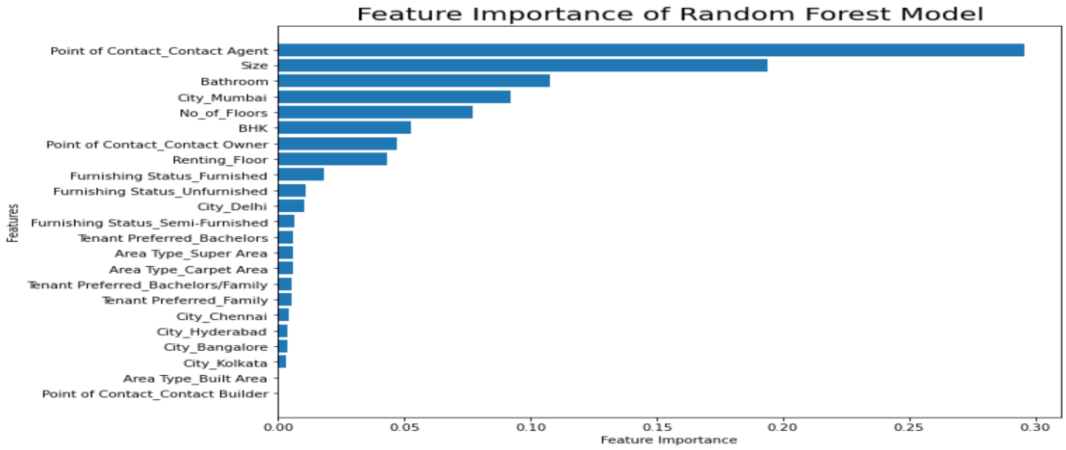
6. Modeling

As the problem addressed involves supervised learning with continuous target variable, the following models are selected for this project; they are Linear Regression, Decision Tree, Random Forest Regression, and Least Absolute Shrinkage and Selection Operator (Lasso) Regression. All the four models are created and fit with the X_train and y_train datasets respectively, predictions are created for the test dataset to evaluate the model's performance. For the evaluation of the models R Squared and RMSE (Root Mean Squared Error) are considered. The selected metrics are calculated for each model respectively.

7. Evaluation

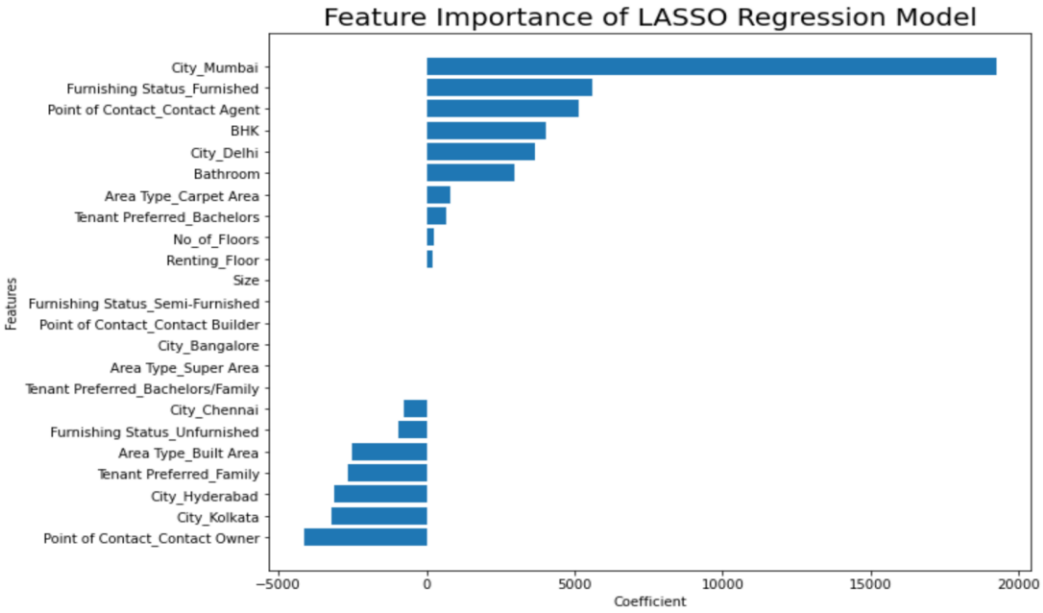
- The evaluation metrics R Squared and RMSE (Root Mean Squared Error) are considered and are calculated for each model respectively.
- The Random Forest Model performed good with low RMSE (Root Mean Squared Error) value as 9770.76, R Squared value is good on the train dataset i.e. 0.9639, but the R Squared value on the test dataset is dropped to 0.76565.This indicates that the model is not generalizing properly and indicates over fitting of the train dataset.

Figure 7: Feature Importance of Random Forest Model



- The Decision tree model can be neglected as the RMSE and the R Squared values are more and the R squared values for the test and train data are having huge difference, which indicates the model is being over fit with the train dataset.
- When observed the R Squared values and the RMSE values of the train dataset and the test dataset of Linear Regression model and LASSO Regression models, the values are almost the same for both the models which are very close. These models are not having the over fitting problem.

Figure 8: Feature Importance of LASSO Regression Model.



- The following figure summarizes the evaluation metrics of all the four models.

Figure 9: Evaluation metrics results summary.

	Linear Regression Model	Decision Tree Model	Random Forest Model	Lasso Regression Model
Model	Linear Regression	Decision Tree	Random Forest	LASSO Regression
R Squared(test)	0.712764	0.626338	0.766566	0.712908
R Squared(train)	0.700134	0.997924	0.964429	0.700121
RMSE	10817.24843	12337.773491	9751.685436	10814.529152

8. Deployment

Though the Random Forest performed well the model is not ready for deployment; however the features that are identified can be useful for the rental sites to concentrate on the identified area which can increase the usage of the site by the people in search of a rental house.

In future the project recommendations are; the project can be improved by applying the PCA (Principal Component Analysis), hyper parameter tuning on the models selected to reduce the problem of over fitting in the model and improve the performance of the models. Cross-Validation technique can be used to assess how well a model performs on a new independent dataset. The simplest example of cross-validation is when you split your data into three groups: training data, validation data, and testing data, where you see the training data to build the model, the validation data to tune the hyper parameters, and the testing data to evaluate your final model.

9. Conclusion

There are several factors that play an important role in the prediction of the house rents in any place. With the above analysis the Random Forest regression model best fits for the analysis of the House Rent dataset selected. As per the above analysis of the House Rent Dataset of India, we can say that in India the houses rents are mostly dependent upon the cities where the houses are situated, the number of Bedrooms, Hall and Kitchen available for houses, Bathroom available and Size of the house. This analysis will be helpful for the rental listing sites for contacting the owners of the rental properties available and updating the sites for easy access for the persons searching for rental houses, thereby increasing the business.

References

Google, Data Science Process Alliance, What is CRISP –DM?, https://www.datascience-pm.com/crisp-dm-2/

Google, medium, Understanding CRISP-DM and its importance in Data Science projects, Zipporah Luna, Jul 20, 2021, https://medium.com/analytics-vidhya/understanding-crisp-dm-and-its-importance-in-data-science-projects-91c8742c9f9b

Google, https://www.researchgate.net/publication/304153916_India's_Residential_Rental_Housing

Google, https://www.businesswire.com/news/home/20211027005747/en/India-Rental-Housing-Markets-Competition-Forecast-Opportunities-FY2027---ResearchAndMarkets.com

Google, inria.github.io, scikitlearn, feature selectionhttps://inria.github.io/scikit-learn-mooc/python_scripts/dev_features_importance.html

Google, towardsdatascience , feature section in machine learning using lasso regression, https://towardsdatascience.com/feature-selection-in-machine-learning-using-lasso-regression-7809c7c2771a

Google, Machine Learning Mastery, How to calculate Feature Importance with Python,*https://machinelearningmastery.com/calculate-feature-importance-with-python/*

Google, towardsdatascience, hyperparameter tuning in linear regression, *https://medium.com/analytics-vidhya/hyperparameter-tuning-in-linear-regression-e0e0f1f968a1*

Google, A Case Study of Evaluating Job Readiness with Data Mining Tools and

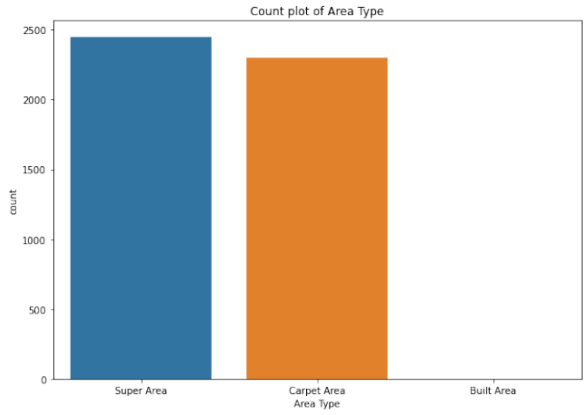
CRISP-DM Methodology, https://infonomics-society.org/wp-content/uploads/iji/published-papers/volume-8-2015/A-Case-Study-of-Evaluating-Job-Readiness-with-Data-Mining-Tools-and-CRISP-DM-Methodology.pdf

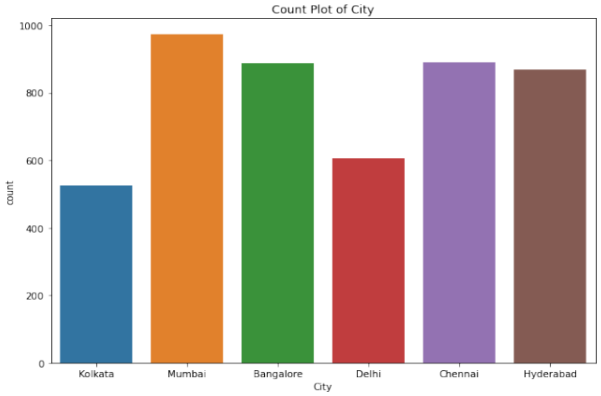
Google, medium.com, PAIRPLOT VISUALIZATION, Sarath SL, Sep 29, 2019.https://medium.com/analytics-vidhya/pairplot-visualization-16325cd725e6#:~:text=Pair%20plot%20is%20used%20to,separation%20in%20our%20data%2Dset.

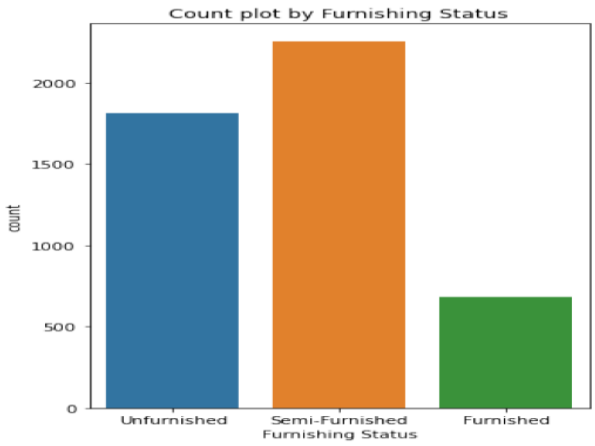
Final Project

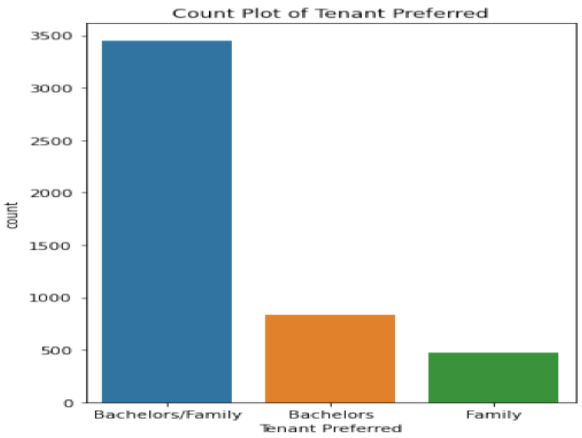
Appendix - A

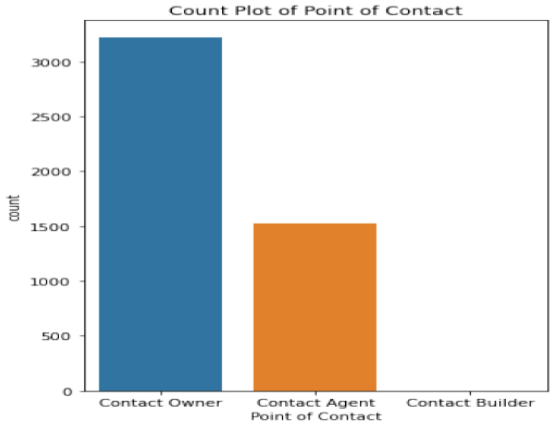
Figure 10: Count plots of the Categorical variables







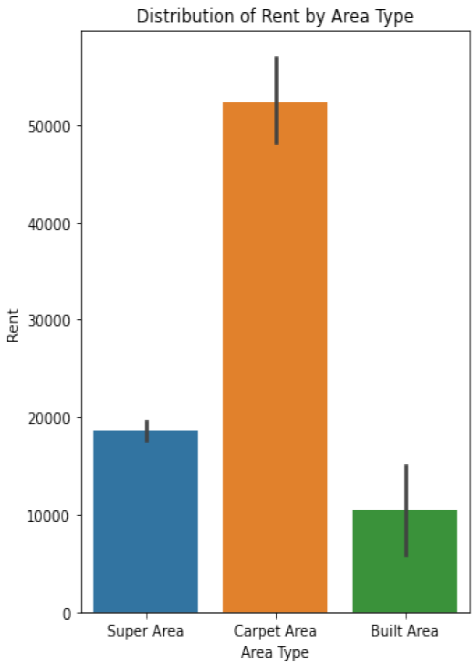


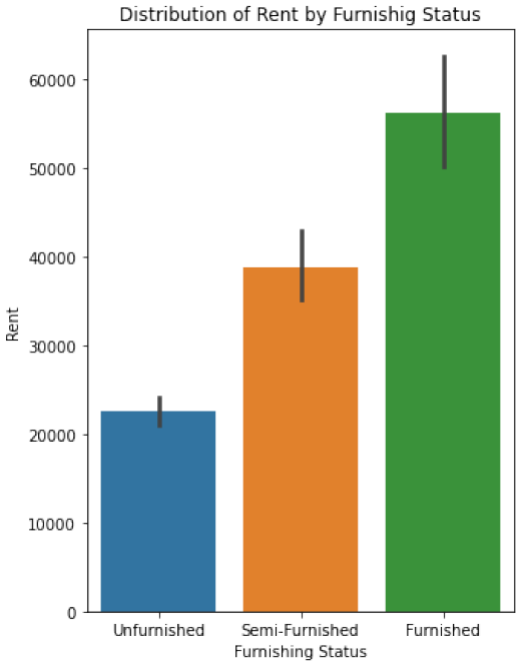


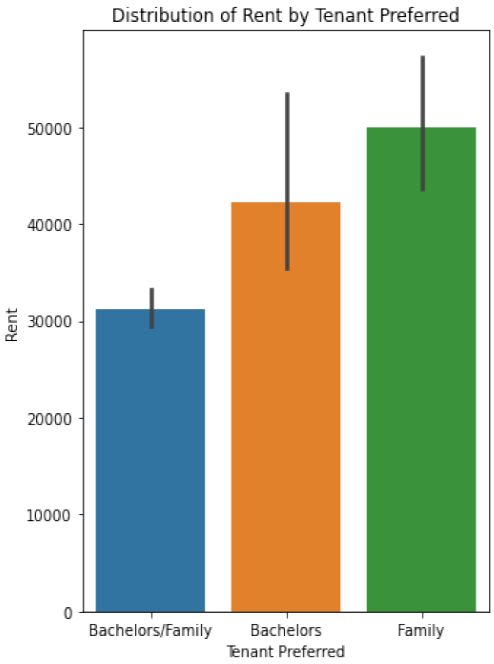
Final Project

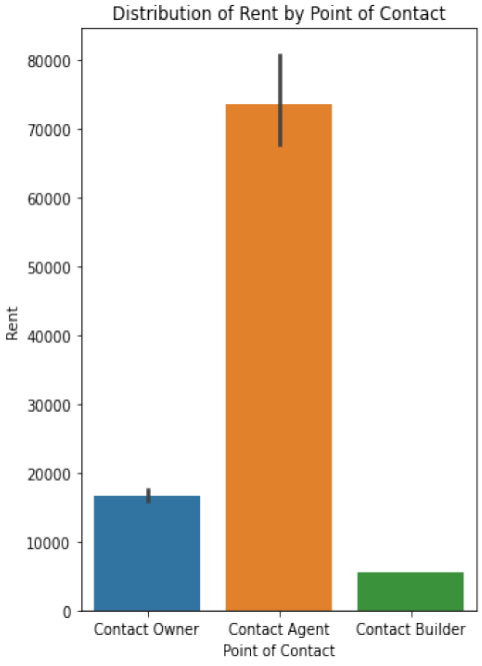
Appendix – B

Figure 11: Distribution of 'Rent' with the Categorical variables.









Final Project

Appendix – C

Figure 12: Box plot of the target variable Rent showing the outliers.

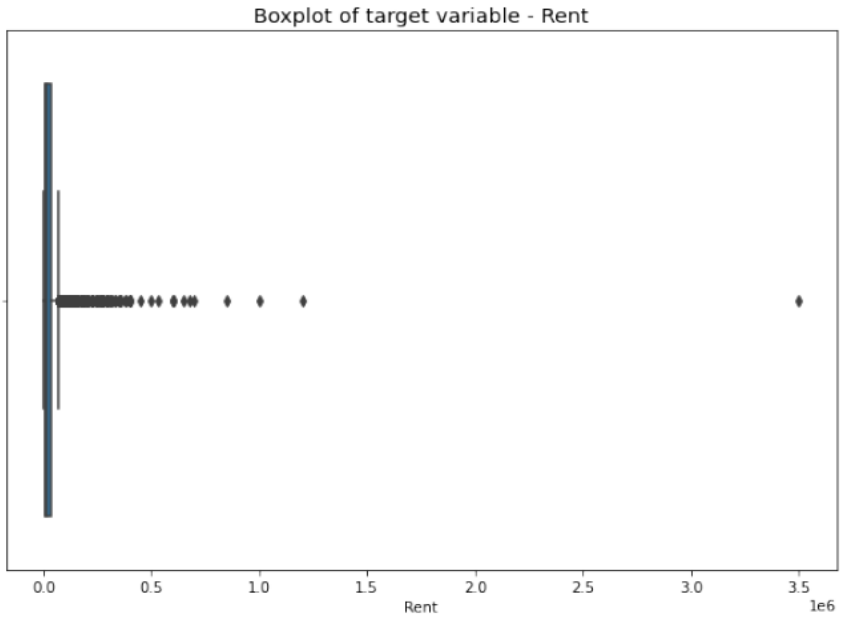


Figure 13: Boxplot of the target variable after removing the outliers.