CS 59000 – Data Analytics in Business using R

Fall 2022

Group 9

House Rental Prediction System

*Likhita Budhavaram[a], Vedhan reddy Gaddam[a], Prem Kumar Ravva[a], Ruthvik Reddy Vanga[a]

[a] Department of Computer Science, Purdue University, Fort Wayne

Fort Wayne, Indiana, 46805

Under the guidance of

Professor Adolfo Coronado
Chair of Computer Science and associate Professor

Department of Computer Science, Purdue University, Fort Wayne

Fort Wayne, Indiana, 46805

Abstract

The real estate industry has expanded in recent years, providing many benefits. With extensive advertising on social media, electronic platforms bombard people with things to purchase by the number of bedrooms and availability of different amenities. Real estate is one of the most critical assets in today's world, especially in a city with potential job hubs that attract a lot of individuals. The primary goal of this project is to estimate the housing market's current price. While doing so, considerations like the number of bedrooms and the availability of other amenities are made. This prediction is meant to assist a customer in finding more practical solutions and better meet their needs. We employed a linear regression model to estimate the cost of the various houses in question. This technique saves the customer from having to speak with a broker, which is another benefit.

Introduction

Today, real estate symbolizes a person's wealth and status and meets one of a man's essential needs. Real estate investments appear to be generally profitable since their property prices don't drop quickly. Changes in real estate prices may impact various household investors, bankers, policymakers, and others. The real estate industry appears to be a desirable place to invest. As a result, estimating the value of immovable property is a crucial economic indicator. Different prediction methods were tested in this study to determine the best-predicted outcomes in calculating a house's selling price compared to the actual price.

This project presents research on regression techniques like linear regression that can be applied to home prediction. Since the original house price predictions were difficult, the best strategy is needed to obtain an accurate projection. Missing features are a challenging component to address in models, let alone a house prediction model, and data quality plays a vital role in predicting house prices. In this manner, feature engineering becomes a significant strategy for making models that will give better precision.

Property values typically rise over time, so it is necessary to determine their current value. This valued value is necessary for property sale and marketability. Qualified evaluators establish these valued values. The drawback of this method is that these appraisers could be partly due to the interests that buyers, sellers, or mortgages have conferred. Therefore, we require an automated prediction model to forecast property values accurately. First-time purchasers and less experienced clients can use this automated technique to determine whether property rates are overvalued or undervalued.

Data Preparation

1. Data Collection

We will go over how to structure the dataset to be accurate for our model, sensible, and well-organized. From the perspective of property price prediction, the steps outlined in this section are crucial.

2. Data Gathering

Data collection is one of the first steps in creating any model. The accuracy of the predictions made by the model depends directly on the precision of the collected data.

In this study, we aim to develop a model that would predict real estate prices using a linear regression technique. We used the Kaggle dataset for housing prices.
The data has about 4746 houses, Apartments, and flats available for rent. It is divided into 12 columns and includes information on BHK, rent, size, number of floors, area, type, locality, city, furnishing status, preferred tenant type, number of bathrooms, and point of contact. The columns in the data set describe the following:

1. BHK: The column describes the number of Bedrooms, Hall, and Kitchen.
2. Rent: The column describes the rent on Houses/Apartments/Flats.
3. Size: The column describes the size of the Houses/Apartments/Flats in Square Feet.
4. Floor: The column describes which floor the Houses/Apartments/Flats are present on with the Number of Floors in total.
5. Area Type: The column describes the size of the Houses/Apartments/Flats calculated on either Super Area, Carpet Area, or Build Area.
6. Area Locality: The column describes the Locality of the Houses/Apartments/Flats.
7. City: The column mentions the city where the Houses/Apartments/Flats are Located.
8. Furnishing Status: The column describes the furnishing Status of the Houses/Apartments/Flats, whether it is Furnished, Semi-Furnished, or Unfurnished.
9. Tenant Preferred: The column describes the Type of Tenant Preferred by the Owner or Agent.
10. Bathroom: The column describes the number of bathrooms in the Houses/Apartments/Flats.
11. Point of Contact: The column provides information about the people to contact for more details regarding the Houses/Apartments/Flats.

Data Cleaning

The data we gather is typically noisy. It could have blank fields, inaccurate data, and outliers. This type of data may negatively impact the accuracy of the predictions made by the model. As a result, it is crucial to eliminate any such noisy data.
The dataset must first be checked for any missing fields. Figure (1) shows that we removed all the entries from our dataset that had empty fields.

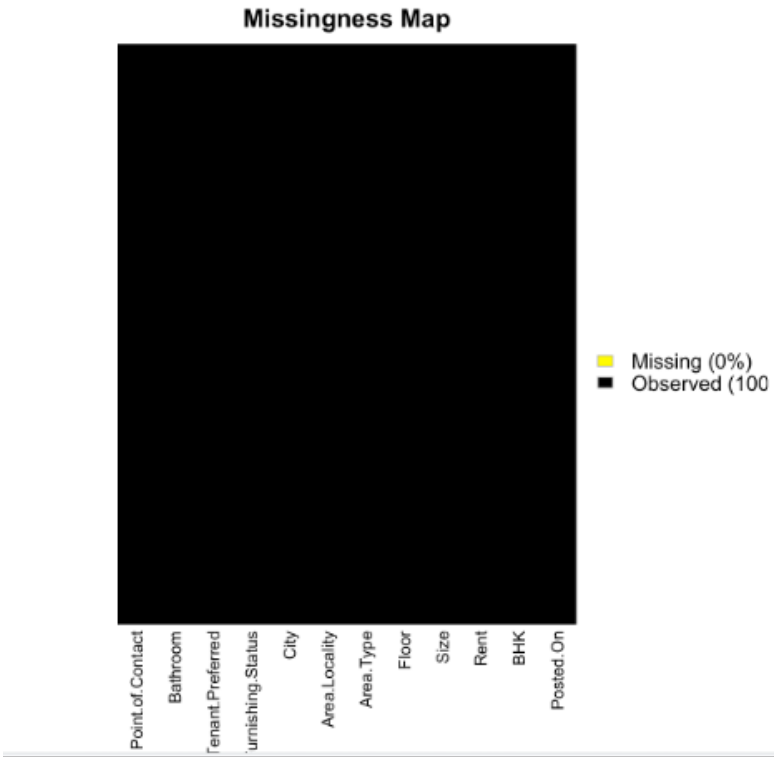


Figure 1: Missingness Map

Checking for outliers is the last stage. Outliers are observations that differ from the rest. The prices of the properties were examined, and if any price differed noticeably from those of other properties in that area, it was removed from the table. After the data cleaning stage, the distribution of real estate prices looks like the distribution plot in figure 1 above.

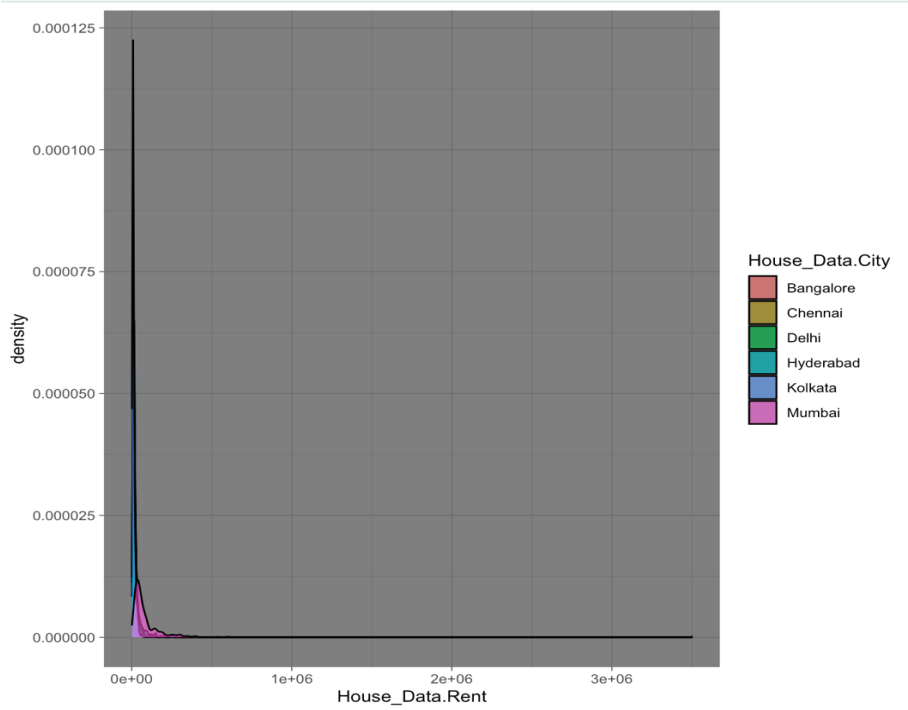


Figure 2. Density distribution of property price

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a type of analysis that seeks out broad trends in the data. These patterns include anomalies and potentially unexpected aspects of the data. Any data analysis should begin with an EDA. Designing statistical analyses that produce relevant findings requires an understanding of the locations of outliers and the relationships between variables. Prices of the houses are probably impacted by a variety of factors, according to location, floor, size, etc. Therefore, basic explorations of stressor correlations are crucial. EDA can offer perceptions of potential reasons that ought to be considered in a causal analysis.

Case 1: The relationship between the city and the rent can be observed in the following bar graph. From the graph analysis, we can see that the rent of the houses is more in Mumbai than in the other places and is least in Kolkata.

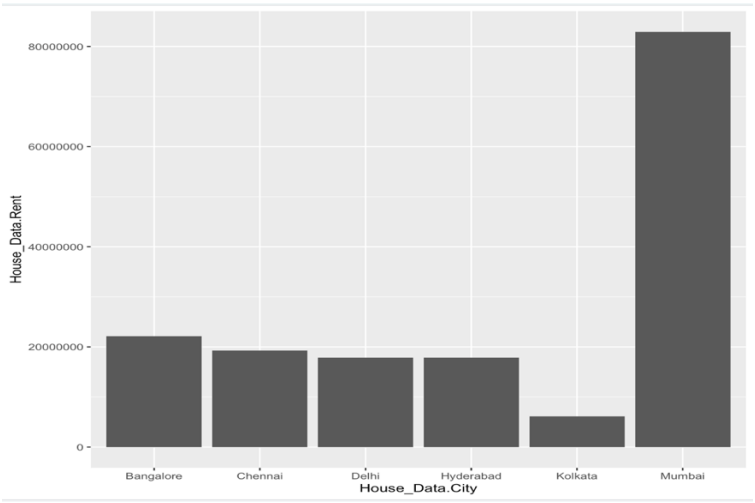


Figure: 3 Distribution of City and Rent

Case 2: The plot depicts the number of houses available for rent in various cities. We can observe from the graph that the number of cities available for rent is more in the Mumbai than in the other places.

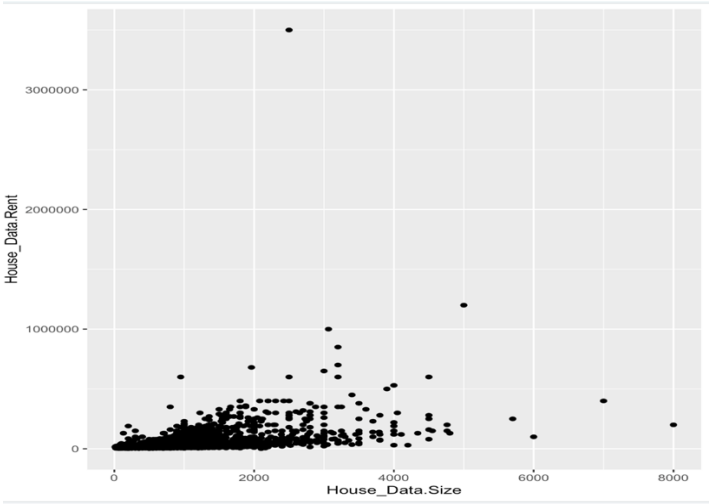


Figure 4: Distribution of Size and Rent

Case 3: This plot depicts the distribution levels of the furnishing levels. The furnishing levels have three stages furnished, semi furnished, and unfurnished. We will see how this will affect the rent of the houses.

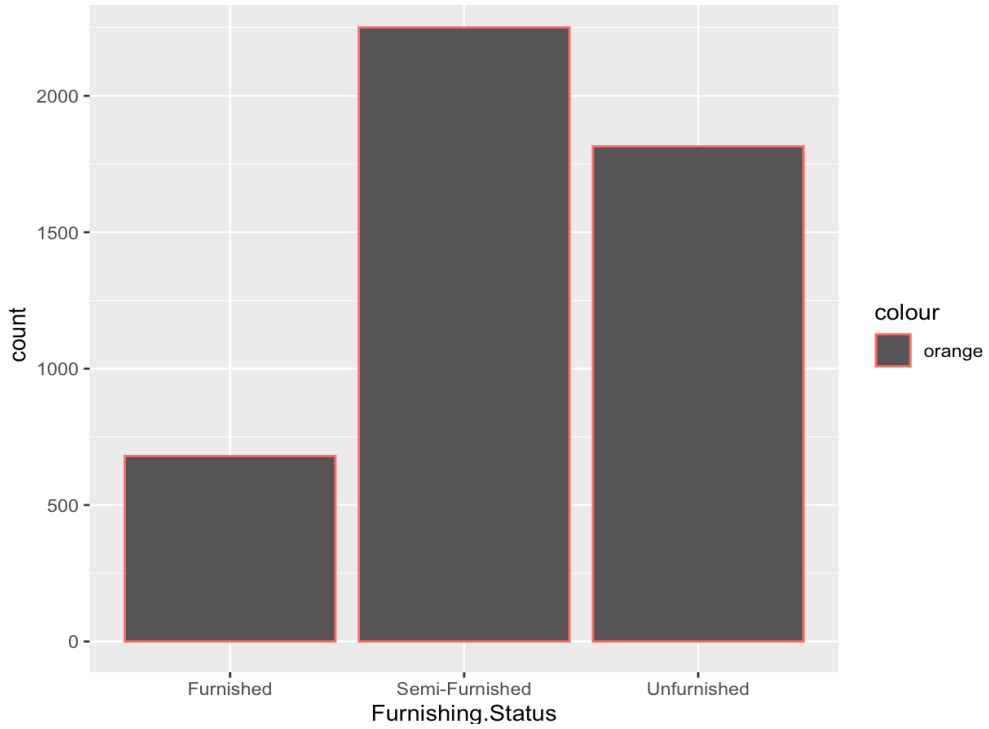


Fig 5: Furnishing status

Using these patterns and comparisons we will use various predictive models to find the price house.

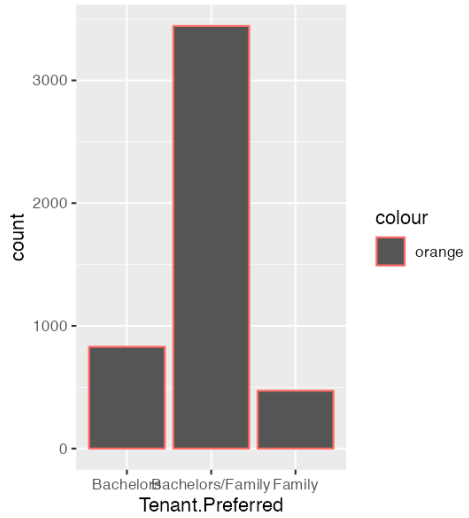


Fig 6: Tenant Preferred

The bar graph shows the patterns of how the variable tenant preferred is distributed along the dataset. It shows that people are neutral about their preference for the tenants they stay with.

Feature Engineering

Feature Engineering, extracts features from unprocessed data. It aids in better communicating a fundamental issue to predictive models, increasing the model's accuracy for unobserved data. The feature engineering method chooses the most practical predictor variables for the model, which is composed of predictor variables and an outcome variable.
The dependent variable is transformed using logarithms in the example below (Rent).
Each graph displays the Rent distribution.

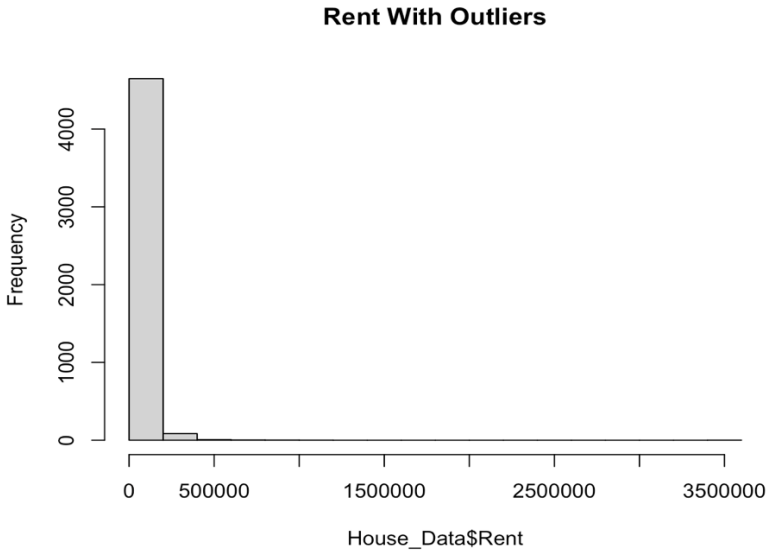


Fig 6: Rent with outliers

The above graph has y and x axis plotted. X axis has the rent and y axois shows its distribution. The above graph shown is different from others because Rent has extreme outliers.

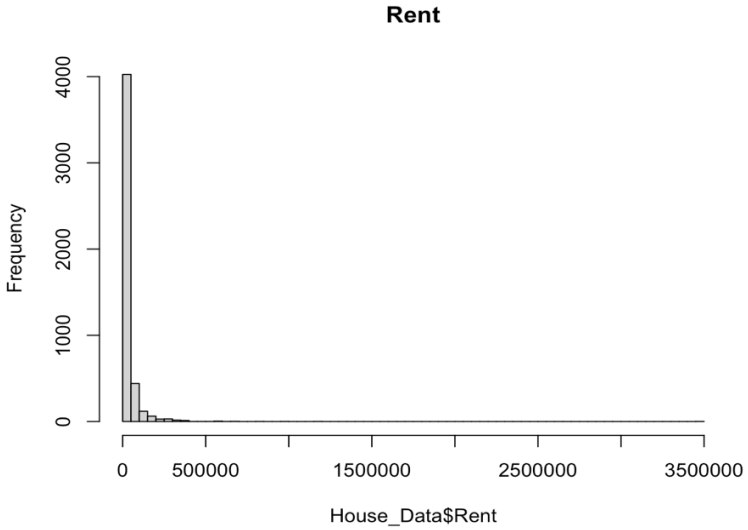


Fig 7: Rent without outliers

The fig 7 shows the data's original distribution without any outliers, and it is evidently right skewed. Which means that the distribution of the data is towards the left.

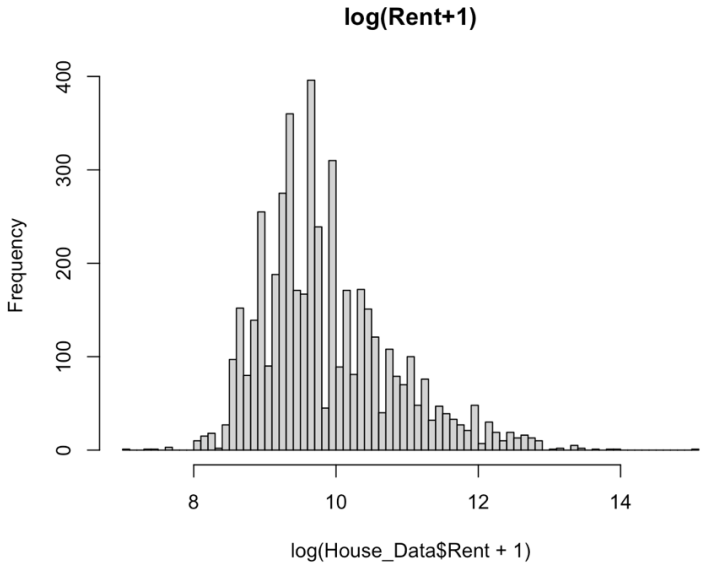


Fig 8: logarithmic of Rent

After applying a logarithmic adjustment to Rent, the result has a shape that is very similar to the normal distribution. The logarithmic transformation is crucial for model strength in this investigation.

Methodology

The main objectives of our project are to develop a system that will benefit customers or users searching for households and make renting houses more efficient. The project aims to provide a virtual system for customers to find a perfect match of houses based on their preferences and customize their needed specifications.

With the improvement in technology, the house rental prediction model bridges the gap between online and offline real estate aspects. The housing sector remains vigilant to face the challenges and provides new strategies that facilitate easy management of rental houses. Hence there is a need to develop a House rental Project. The house Rental Prediction model is based on the Apartment House for rent in various locations. It also focuses on various areas like:

1. How does the rent vary depending on the size of the house?

2. How are other parameters affecting the price of the house?

3. Average house rent by the city?

We have seen from the plots that furnished status has a direct effect on the house rent. Furnished status has three levels of values like furnished, semi-furnished, and unfurnished. Depending on these values the rent is predicted. Other parameters like the size of the house, number of rooms, and number of bathrooms, can also make a significant impact on the prediction of the rent of the house. We have calculated the average rent of the house using

the mean and plotted it to find the distribution of it in the various cities. We will discuss the algorithm we have used for the prediction of the property price and how the training and testing of the model will take place.

Correlation Plot

A tool for analysis that combines correlation coefficients between an x-axis and a y-axis where many variables are found is the correlation matrix or correlation table. There are three possible outcomes from the correlation matrix.

- A positive correlation indicates that two variables or elements are linked because they move in the same direction.
- A lack of connection exists between the two variables, called a neutral correlation.
- The two variables move in the opposite direction when there is a negative connection.

Here from the dataset, we can see that.

- Checking the correlation between our variables, here we can see that the Bathroom, BHK, and Size have a very high correlation with the rent variable, and floor and locality have less correlation with the variable.

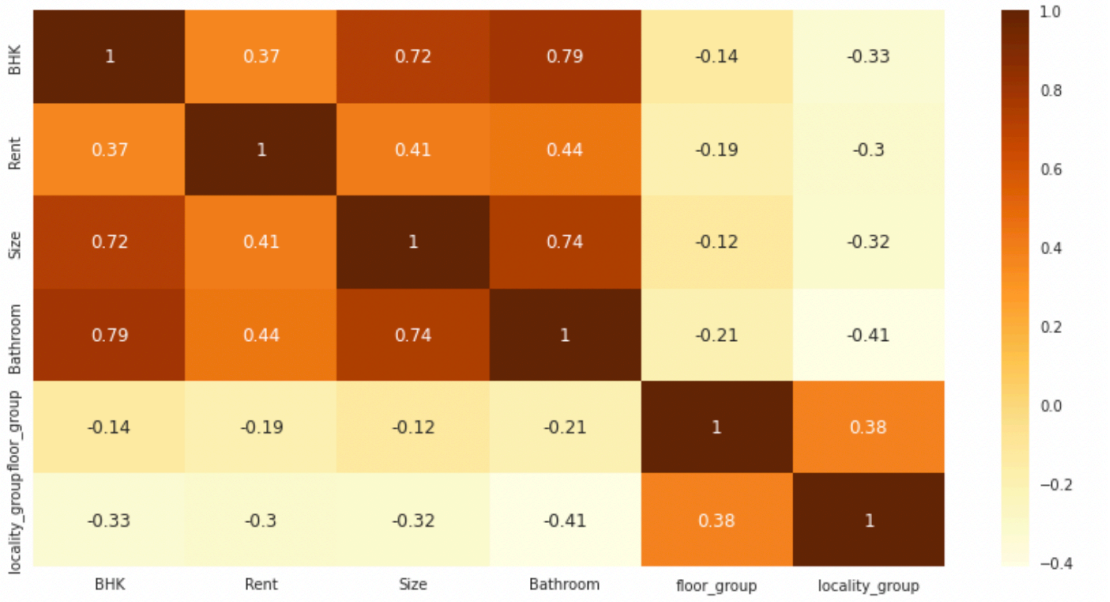


Fig 6: Correlation Plot

Modeling

Now that the data is cleansed of noisy data and preprocessed, we can finally use it for prediction. On this page, In this section, we'll talk about the algorithm we utilized for the training and testing of the model, and the price of the property is predicted. The development, training, and application of machine learning algorithms that simulate logical decision-making based on accessible facts are known as AI modeling. Advanced intelligence approaches including real-time analytics, predictive analytics, and augmented analytics are supported by AI models, which act as a foundation.

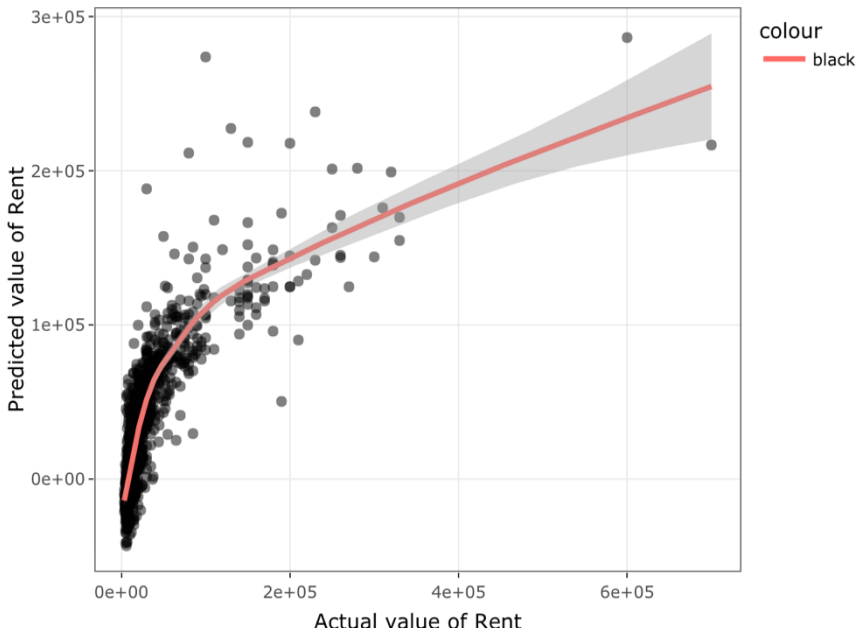
Training and Testing Data

The data used to train an algorithm or machine learning model to anticipate the outcome that your model was designed to predict is known as training data. You need unknown data to test your machine learning model after it has been constructed (using your training data). You can use this data, which is referred to as testing data, to assess the effectiveness and development of your algorithms' training and to modify or optimize them for better outcomes.

The data set is split into two categories: training data and test data and data testing. In this case, 20% of the data is used for testing and the model is trained using 80 of the remaining data. In order to fit the training data into the linear model after dividing the data, the regression model gets the indicated line of regression, knn, and random forest.

Linear Regression

Regression modeling is one of the most fundamental of the many data mining methods that have been created. Simply creating a mathematical model from measured data is regression modeling. According to this approach, an output value can be explained by a set of input values.



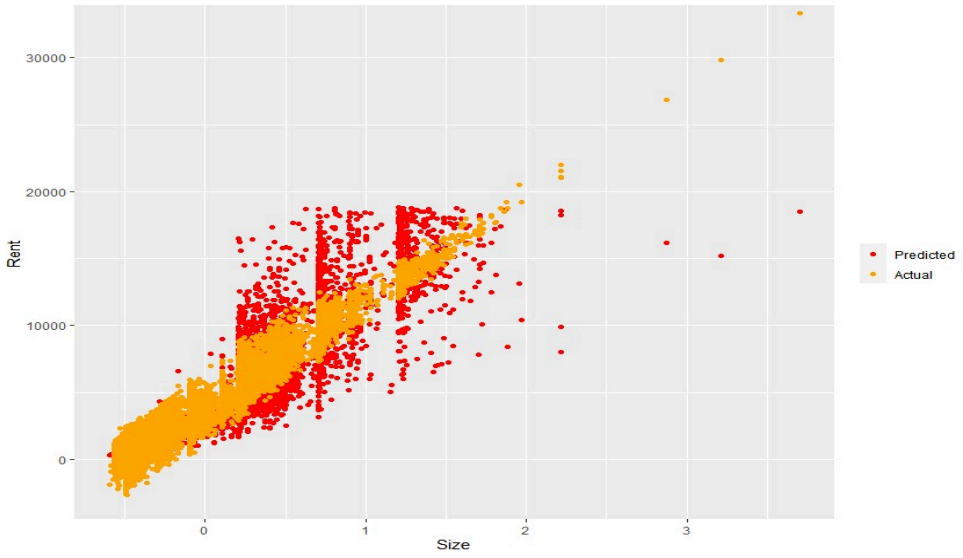


Fig 7: Linear Regression

Regression modeling with linear assumptions is a particular type of regression modeling. Our model attempts to model a linear relationship between dependent variables (Y) which will be the rent and independent variables (X) which will be the various variables in the dataset. From the correlation plot, we found that the size has the highest correlation and hence we use that as the x variable and predict the model. The above graph is the linear regression model and we are predicting the rent on the variable size.

K- Nearest Neighbors

KNN is a supervised learning method that predicts the output of the data points using a labeled input data set. One of the simplest machine learning methods, it can be used to solve a wide range of issues. It primarily relies on the similarity of features. A data point is classified into the class to which it is most similar using KNN, which examines how similar it is to its neighbor. KNN is a non-parametric model, which implies that it does not make any assumptions about the data set, in contrast to most algorithms. As a result, the algorithm is more efficient because it can handle real-world data. Because KNN is a lazy algorithm, it memorizes the training data set rather than inferring a discriminative function from the training data. KNN can be used to solve problems involving classification and regression. A new data point is classified into the target class using the attributes of its nearby data points via the supervised machine learning method KNN, or K Nearest Neighbor. This means that the new point is assigned a value based on how closely it resembles the points in the training set.

1. First, the distance between the new point and each training point is calculated.
2. The closest k data points are selected (based on the distance).
3. The average of these data points is the final prediction for the new point.

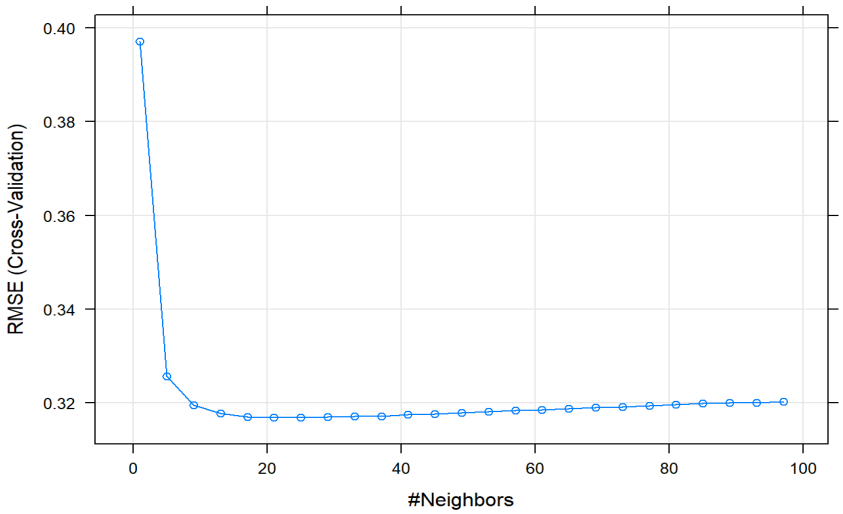


Fig 8: KNN modeling

From the above graph, we can see that on running the KNN model we have an accuracy of around

Random Forest

A well-liked supervised machine learning approach called random forest is utilized to solve both classification and regression issues. Its foundation is the idea of ensemble learning, which enables users to mix various classifiers to solve complicated problems and enhance the performance of the model.

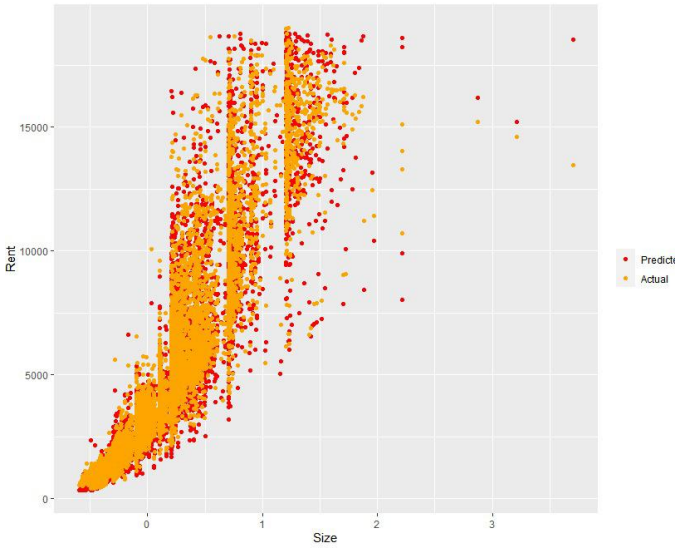


Fig 9: Random Forest

There are many decision trees on which the random forest method is based except for the outcomes of each tree's forecasts. It decides the outcome based on the forecasts that receive the most votes.

Results

After implementing the predictive models such as linear regression, Knn, and random forest we have found the following results. R2 The variance of a dependent variable is measured by the -score. that changes and is predicted by the independent variables from 0% to 100%. If the proportion of the overall variation that is explained by the model to the entire variance is 100%, the two variables are fully correlated, meaning there is no variation at all. Regression models lose more and more of their predictive power as how much R2 their is-score continues going down. R2-score details the number of data points that fall within the produced line using the regression formula. The R2 for our model 0.80 is the -score. Here, it can be shown that our chosen model can explain 80% of the variance of the dependent attribute or variable, whereas the remaining 20% of the random forest model yielded the worst scores, followed by linear regression with 54% and KNN with the least accuracy, respectively. For all error terms, this order is the same. The random forest is the best choice under these circumstances. In conclusion, further feature engineering techniques can be used to enhance the models. Additionally, a higher range of hyperparameters can be adjusted to get a better model for KNN and random forest, and new techniques, like gradient boosting regressor, can be tested.

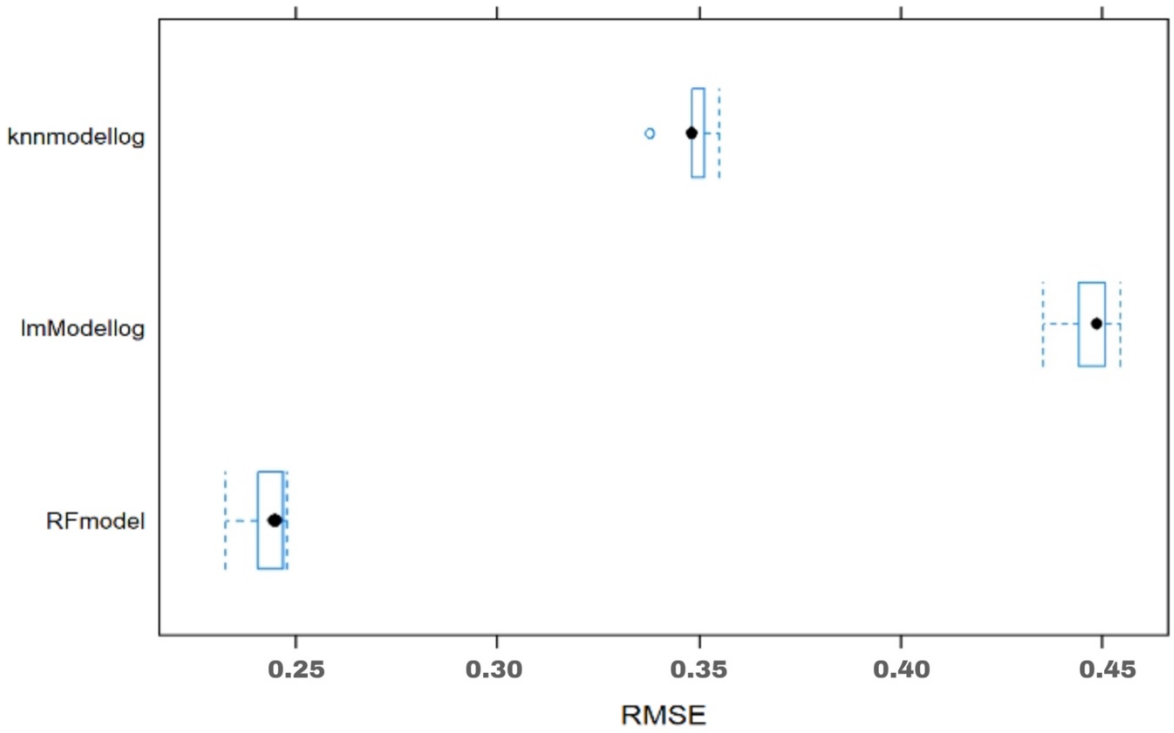


Fig 10: Comparison graph

Dashboard

A dynamic dashboard has a side toolbar that has a dashboard tab and three tabs that have linear Regression, K-Nearest Neighbors, and random forest. A refresh button that can dynamically update the contents of the dashboard is shown in fig 11.

Fig 11: Dynamic Dashboard

1. The dashboard shows the results of the linear regression model.

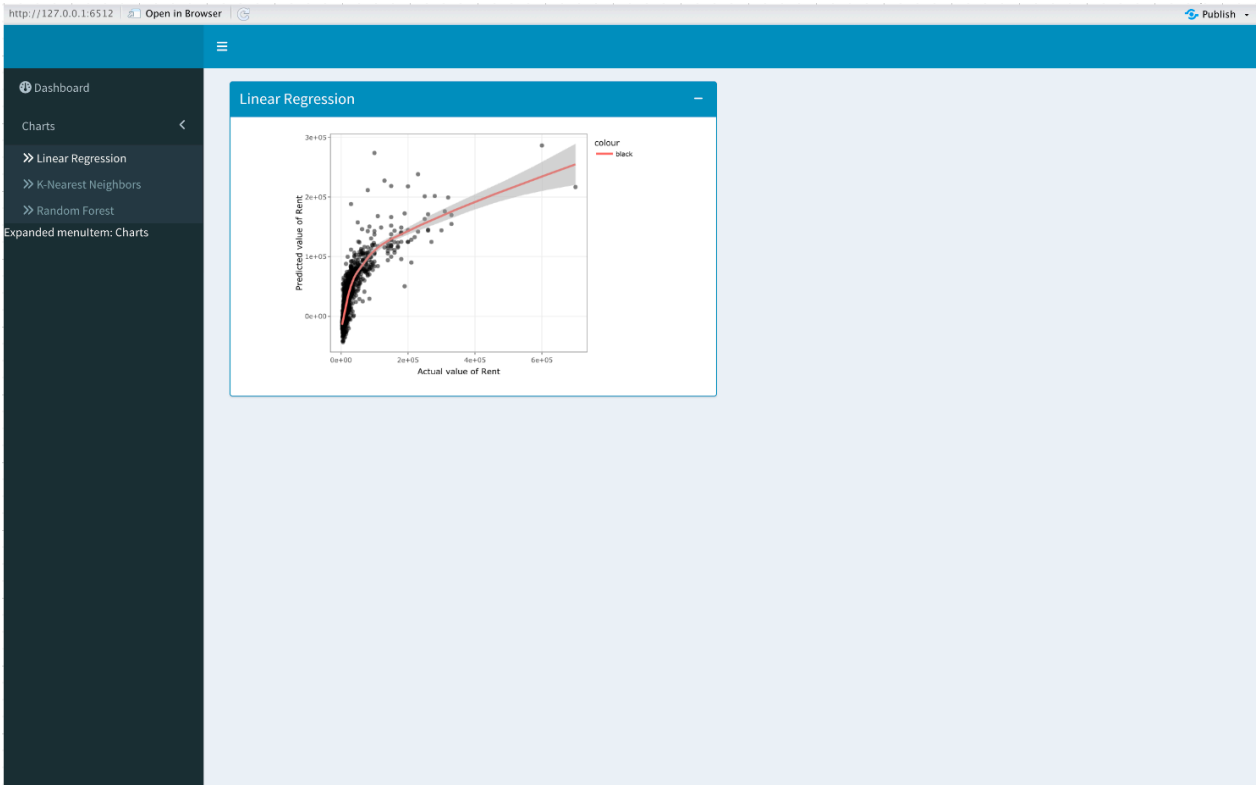


Fig 12: Dashboard tab for Linear Regression

2. The dashboard shows the results of the K-Nearest neighbor model

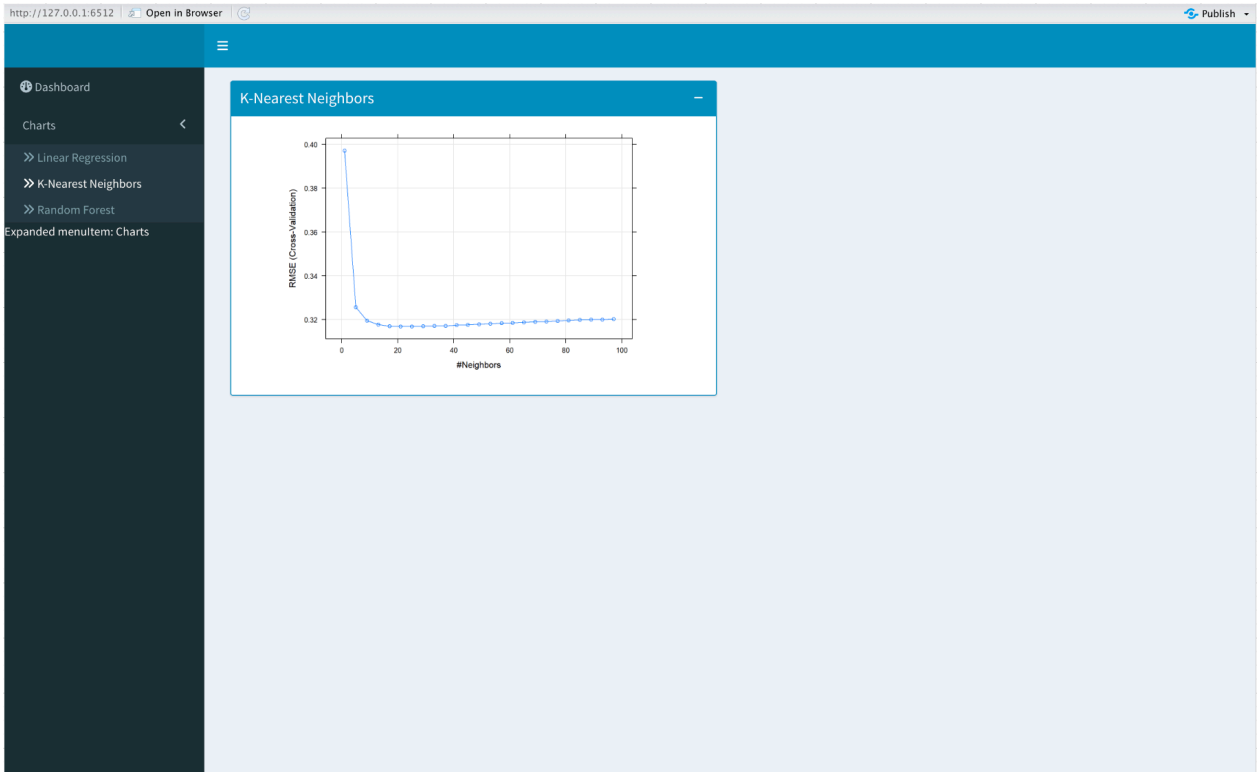


Fig 13: Dashboard tab for KNN

3. The dashboard shows the results of the Random Forest model.

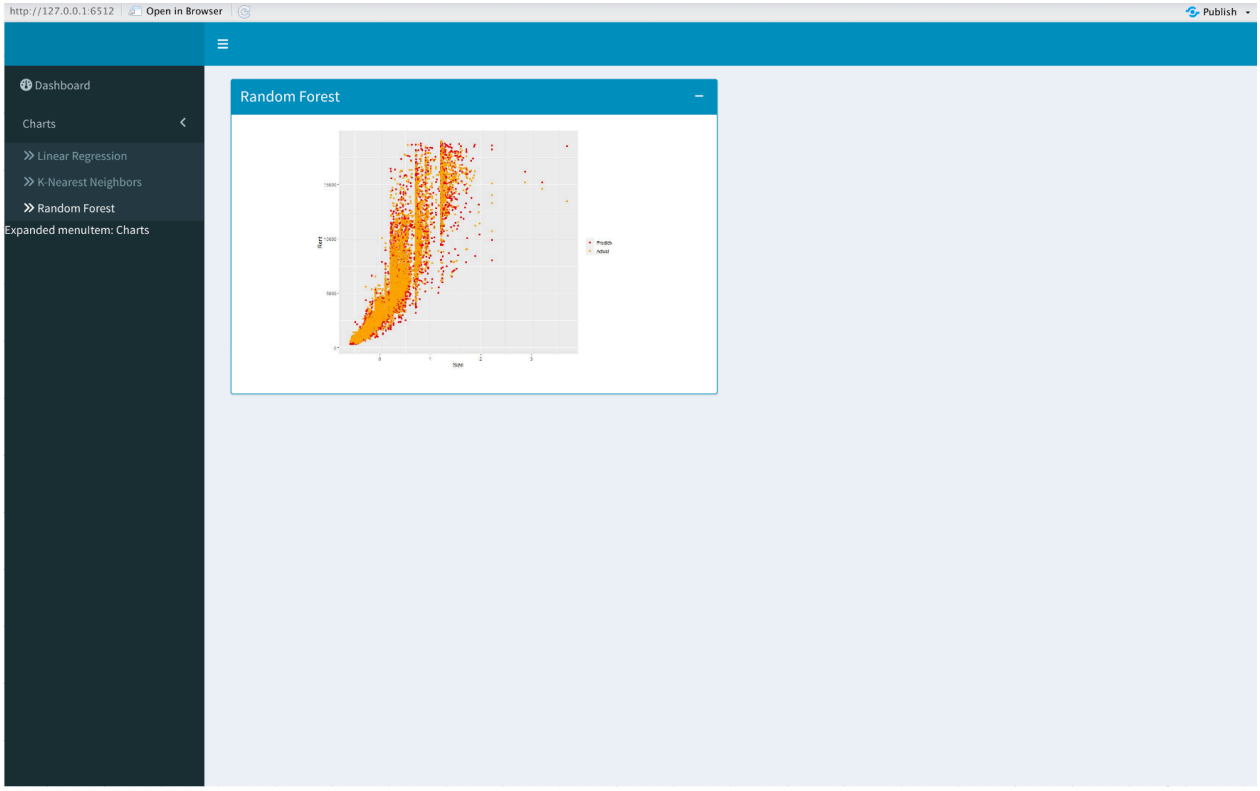


Fig 14: Dashboard tab for Random Forest

Conclusion

In this project, we present a technique for forecasting real estate prices in various locations. We used a dataset from Kaggle with 12 variables for each home, including location, size, security furnishing status, etc. We cleaned up the dataset by removing any erroneous information and outliers. We showed data visualization of various variables. The linear regression method will be used and then fitted to a subset of this dataset, which will then be used to test the model that was chosen. We have also used KNN and random forest for modeling and found that the random forest has more accurate and provided better results than the other models. It is possible to develop this technique further to forecast property prices in other Indian cities and rural regions. Live webpages can also be created using it.

Acknowledgment

The authors would like to thank Professor Adolfo Coronado for helpful discussions on the project on R and predictive models.

References

1. https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset
2. https://shiny.rstudio.com/
3. https://github.com/rstudio/shinydashboard/blob/gh-pages/_apps/sidebar-expanded/app.R