MAT394 Report - Prediction of housing Prices

Group 9

Aravindan T S 1910110080

Adhitya Swaminathan 1910110025

Akshay Jayakumar

Jhananii H

ABSTRACT

This model predicts prices of houses based on features like the square footage of the house, the number of bedrooms, the number of floors, etc. We have used two models. A multiple linear regression model and a random forest model to predict the prices. We then compare the models to see which one yields the better result. The datasets were obtained from Kaggle.

INTRODUCTION

The aim of our model is to predict housing prices based on the number of bedrooms, number of bathrooms, square footage of the lot, square footage of the house, number of floors, condition, square footage of the basement and square footage of the house excluding the basement. We use two models to predict the housing prices - a multiple linear regression model and a random forest model. The random forest model uses various decision trees to predict the price and in multiple linear regression, the price is predicted as a linear combination of every other attribute.

METHODOLOGY

We start by importing the following libraries:

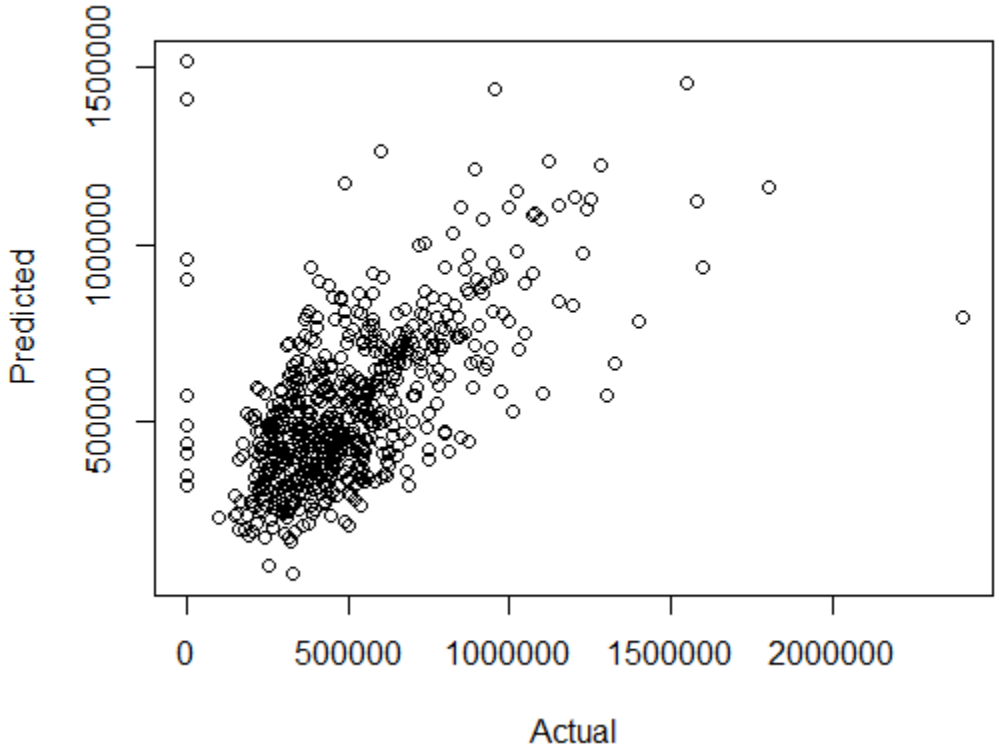
1. Dplyr - Used for manipulating the data
2. Ggplot2 - Used for data visualization
3. Catools - Used to split the data into training set and test set
4. Corrgram - Used to make a correlation matrix plot
5. randomForest - Used to make a random forest model.

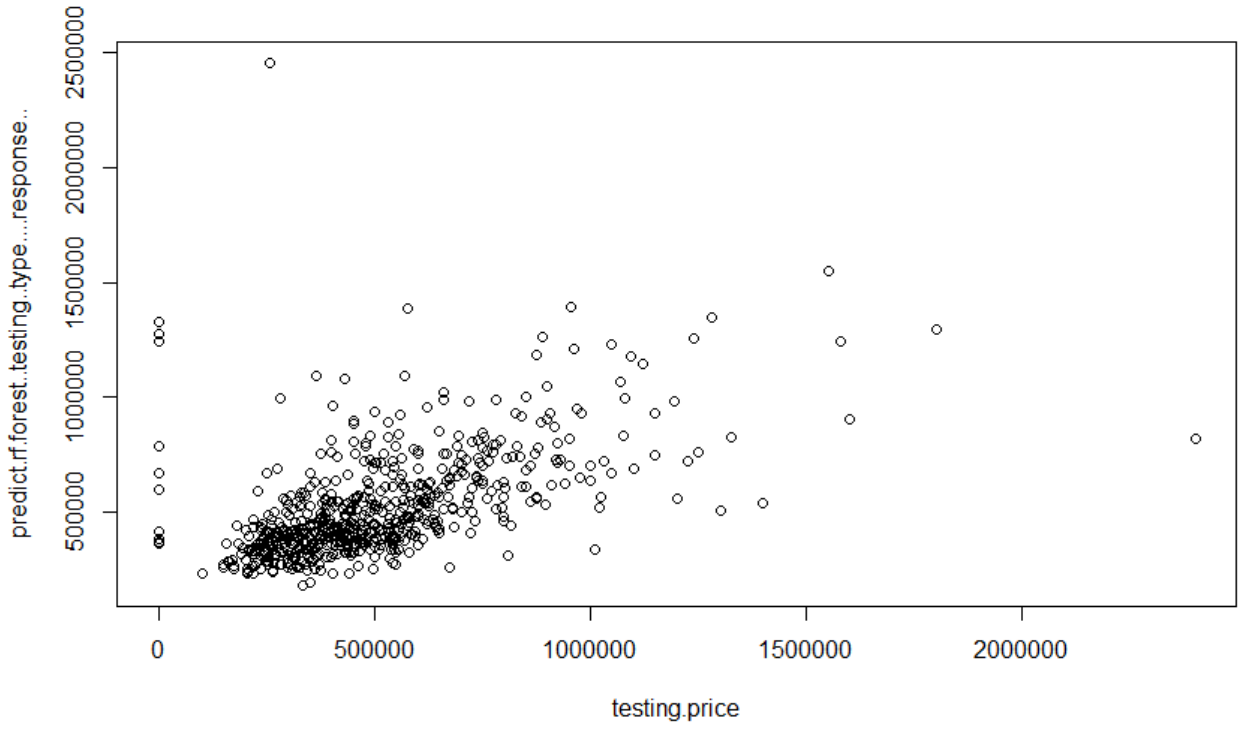
We then import the dataset and start checking for any missing values which in our case were none. Then we use the corrgram library to make a correlation plot to see which are the parameters affecting the price which we are supposed to predict. We then remove parameters like date, waterfront etc which have very negligible effect on the prices. Our data is now cleaned and we can proceed to the next step which is splitting the dataset and creating the necessary models.

We use the caTools library to create a 80-20 split for our dataset. This library randomizes the dataset and splits it into 80 percent which we use to train our data and the rest 20 percent is used to check how accurate our model is. We then proceed to create a multiple regression model and a random forest model and fit our data. Then we use the model on our test set and make two dataframes, one for multiple linear regression and the other for random forest which contain actual and predicted values. Then we plot both graphs and check respective rmse values.

RESULTS

The graphs from both models are attached below.





The rmse value for our random forest model was 169172.5 and 219985.3 for the regression model.

CONCLUSION

We have found the random forest model to be more efficient in the prediction of housing prices than the multiple regression model.

REFERENCES

1. Abdul Qureshi, *Multiple Linear Regression using R to predict housing prices,* https://medium.com/@aqureshi/multiple-linear-regression-using-r-to-predict-housing-prices-c1ba7fe1674a (accessed April 25, 2021).
2. *R - Multiple Regression,* https://www.tutorialspoint.com/r/r_multiple_regression.htm (accessed April 25, 2021).
3. *Random Forest in R | Random Forest Algorithm | Random Forest Tutorial | Machine Learning | Simplilearn,* https://www.youtube.com/watch?v=HeTT73WxKIc (accessed April 25, 2021).