COMPUTATIONAL METHODS FOR ECONOMICS

Predicting the housing prices in household surveys in Uganda

By Kartikay Dhall, Malhaar Khanna, Rishiraj Datta

Summary

In order to estimate the rental value of homes in low-income nations like Tanzania, Uganda, and Malawi, the research article compares the performance of conventional hedonic pricing models versus machine learning methods.

The study developed hedonic pricing models and machine learning models using information from household surveys carried out in these nations. The machine learning models included a variety of input variables, including demographic and socioeconomic traits of tenants, whereas the hedonic pricing models used standard input variables like location, size, and quality of housing.

The study's findings demonstrated that machine learning models performed better at forecasting rental values in all three nations than hedonic pricing models. The predictions made using the machine learning models were more accurate because they could capture the subtler correlations between the input variables and rental values.

The ability of machine learning models to manage large levels of housing market heterogeneity is one of the key benefits of employing them to forecast rental values in low-income nations. This is crucial in many nations since housing conditions and quality can differ greatly even within relatively small geographic areas.

Overall, the research offers insightful information about the potential of machine learning models to enhance our comprehension of the housing market in low-income countries and to guide housing policy decisions.

Introduction

The purpose of this study is to determine whether hedonic pricing and machine learning models can accurately forecast the rental value of homes in low-income nations, particularly Tanzania, Uganda, and Malawi.

The study makes use of a number of variables that are pertinent to forecasting the rental value of homes in order to answer this research issue. These elements are:

1.Location: The geographic location of the home is a factor that can significantly affect how much it rents for. Latitude and longitude coordinates are used in the study to determine location.

2. Size: The size of the house, measured in square metres, is a key component in determining its rental value.

3. Housing quality: The physical state of the home, including its age, construction methods, and general quality, is referred to by this variable.

4. Age, income, level of education, and employment are just a few examples of the tenants' demographic and socioeconomic variables that are included in this variable.

The study develops hedonic pricing models and machine learning models using these characteristics to forecast rental values in the three nations. The machine learning models use a more varied set of input factors, including demographic and socioeconomic features of tenants, whereas the hedonic pricing models use more conventional input variables, such as location, size, and quality of housing.

Methodology

The research predicts the rental value of homes using two machine learning algorithms, Random Forest and Gradient Boosted Tree. These methods are contrasted with conventional hedonic pricing models, such as the Tobit and Ordinary Least Squares (OLS) models.

Both the Random Forest and Gradient Boosted Tree ensemble learning techniques mix different decision trees to get predictions that are more accurate. Due to their capacity for handling complex data and capturing non-linear correlations between variables, these algorithms are frequently utilised in machine learning.

One would need to collect and clean pertinent data from household surveys in Tanzania, Uganda, and Malawi, similar to what was done in the paper, to replicate the ML/econometric exercise outlined in the paper. Then, based on a set of input factors, one would need to construct and train Random Forest and Gradient Boosted Tree models to forecast the rental value of homes.

These models' accuracy can be compared to more established hedonic pricing models like the OLS and Tobit models. It is crucial to remember that the quality and amount of the input variables utilised, as well as the specific parameters and methods employed in the modeling process, will all have an impact on how well the machine learning models perform.

Overall, a thorough understanding of data cleaning, data analysis, and machine learning algorithms would be necessary to replicate the ML/econometric exercise reported in the research utilizing pertinent machine learning approaches.

Results from the analysis

The findings from the econometric and machine learning analyses presented in this study can be used to make predictions about the rental values of homes in Tanzania, Uganda, and Malawi.

First, the study discovered that in terms of prediction accuracy, both machine learning models—Random Forest and Gradient Boosted Tree—performed better than more established hedonic pricing models, such as OLS and Tobit models. This means that using machine learning techniques to estimate rental values in these nations may be more successful.

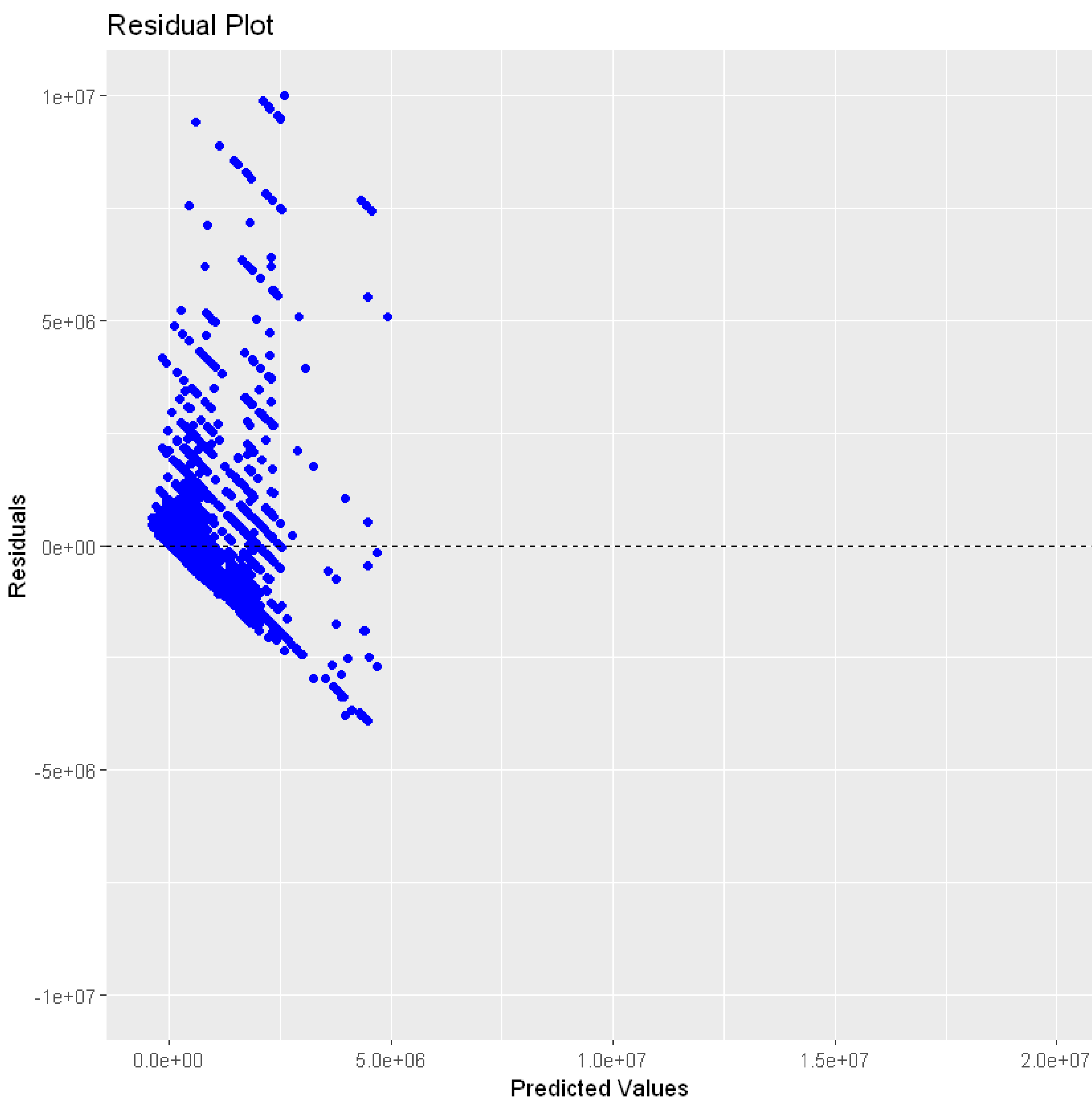
Second, the study discovered that factors like location, size, quality, age, building type, income, and education level were some of the most important predictors of rental value in these nations. These results support earlier studies and offer more proof that these elements have a significant role in determining rental value in the area.

Last but not least, the study emphasises the potential utility of leveraging data from household surveys to forecast rental values, as this strategy may be more practical and affordable than other approaches like property valuation or real estate market analysis. It is crucial to remember that the quality and representativeness of the survey data, as well as other elements like the local political and economic situation, may have an impact on how accurate these predictions are.

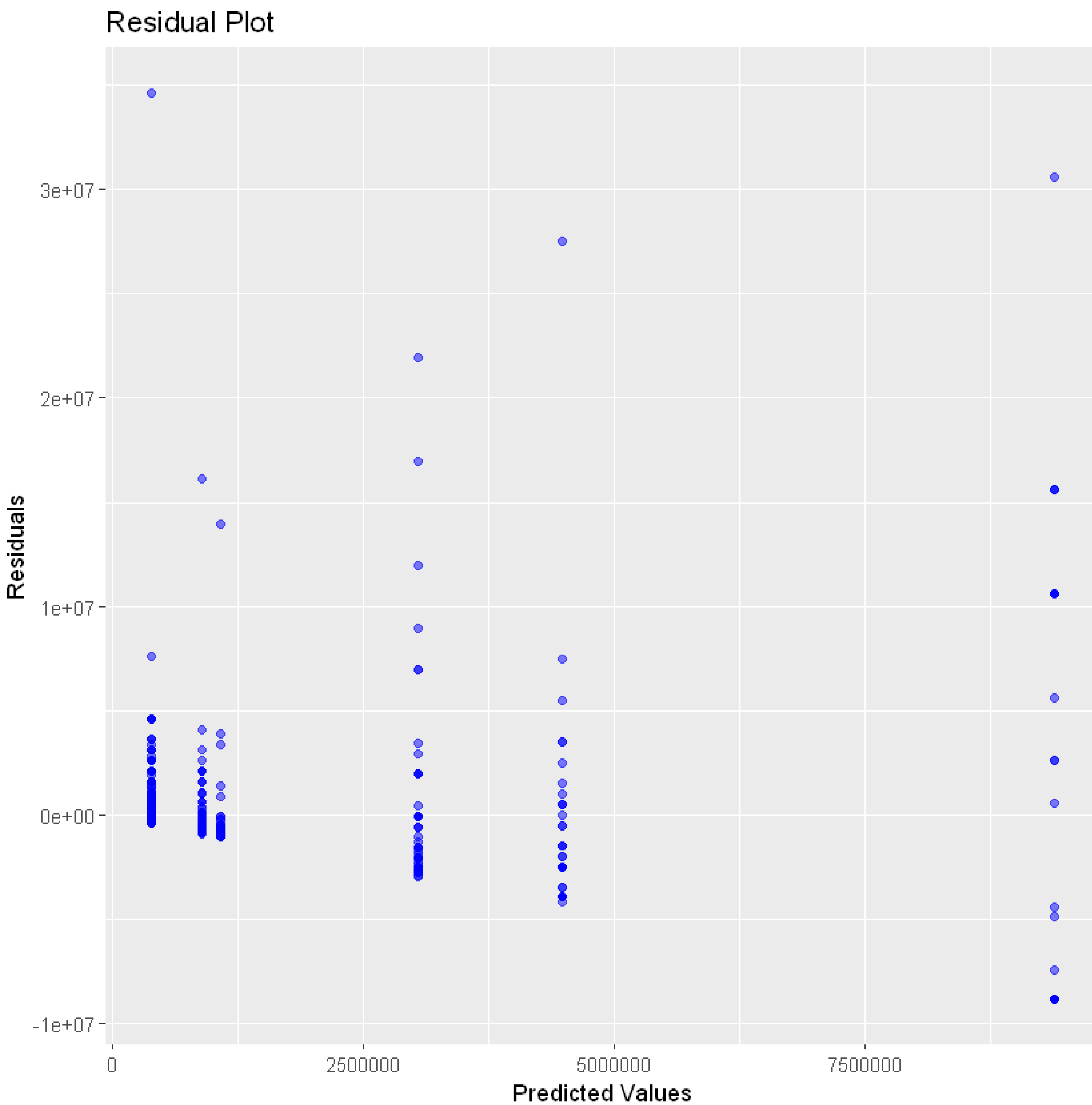
Overall, the findings of this study can help inform policymakers and real estate stakeholders in Tanzania, Uganda, and Malawi on how to better predict and understand rental values, which can ultimately lead to more effective policy and investment decisions.

Graphs

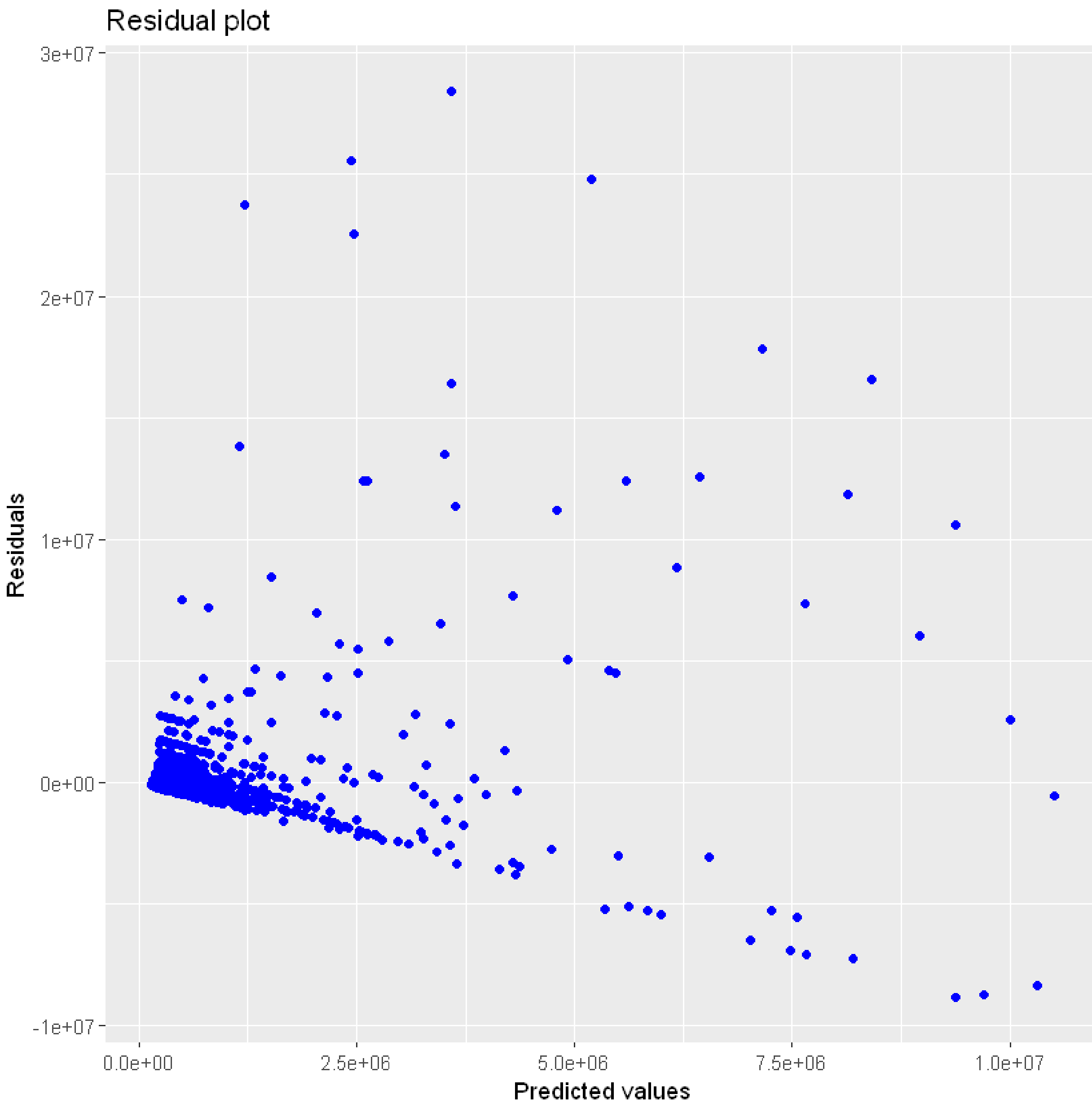
- **Residual Plot for OLS Regression Model**



Tree based Regression Model



Random Forest Regression Model



Gradient Boosting Method

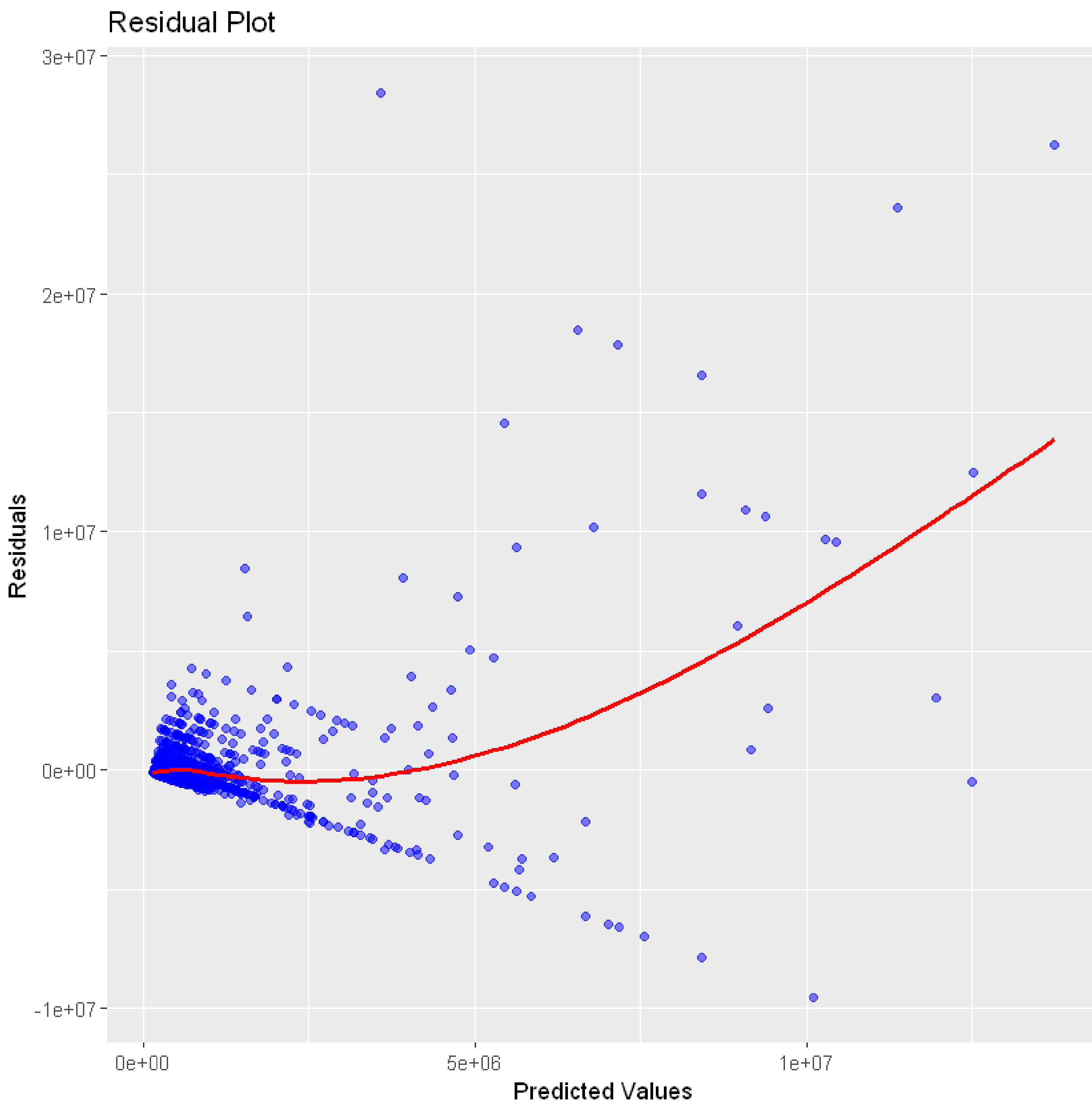


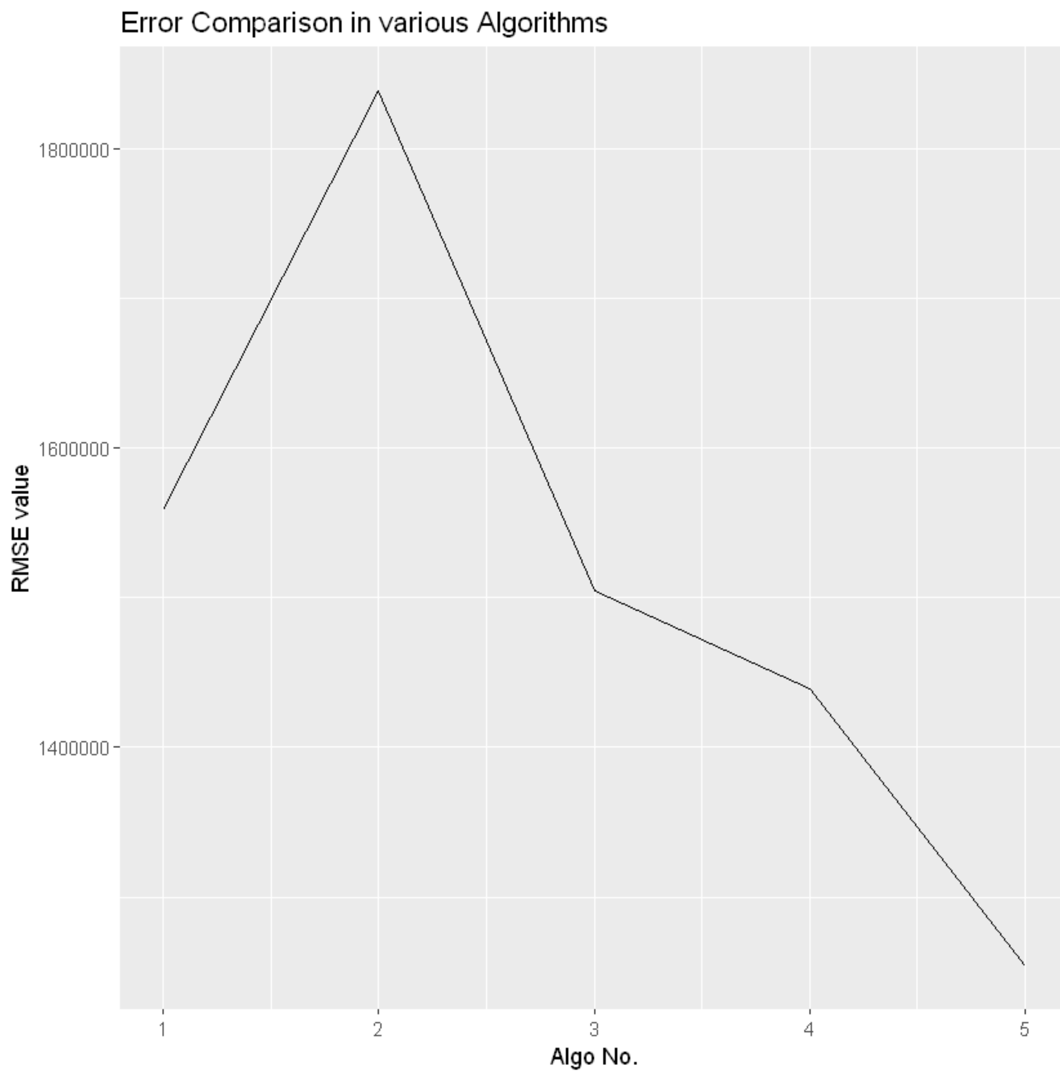
Table for Root Mean Squared Error Values for Various Econometric Methods Used -

Method	RMSE Value	Normalized Error
OLS	1558824	1.00
Ridge	1838892	1.18
Tree	1504464	0.96
Random Forest	1439510	0.92
Gradient Boost	1253958	0.80

It is clearly evident that Gradient Boost Algorithm gives the least error out of all the ML methods we have implemented to build models on our dataset.

	Uganda	
	2010	2012
OLS^{+}	1.00	1.00
Ridge	0.94	0.98
LASSO$^{+\,+}$	0.88	0.96
Tree	0.83	0.86
Bagging	0.78	0.83
Forest	0.82	0.91
Boosting	0.87	0.88

Comparative Error values (as given in the Research Paper)


Error Comparison in various Algorithms
1800000
1600000
1400000
RMSE value
1
2
3
4
5
Algo No.

Interpretation of the coefficients and their relevance in the context of the research paper.

	housing_price	Rent	area_property	cooking_fuel	electricity	electricity_ESCOM	MTLtelephone	source_drinkingWater	toilet_facility	disposal_facility	use_bednet
housing_price	1.00000000	0.145637542	0.014913360	0.200266185	-0.295887837	0.0183385335	-0.12218527 1	-0.244048322	-0.229374800	-0.109166326	-0.0318236301
Rent	0.14563754	1.000000000	0.001198433	0.067299877	-0.117298602	0.0087592067	-0.027982557	-0.096144964	-0.101204239	-0.044588504	-0.0124543597
area_property	0.01491336	0.001198433	1.00000000	-0.001844503	-0.007906939	0.0363187542	0.002756966	0.005288152	0.006577243	-0.002301272	-0.0018163260
cooking_fuel	0.20026619	0.067299877	-0.001844503	1.000000000	-0.512692772	0.0272417699	-0.071386480	-0.494255776	-0.294385356	-0.176811547	-0.0434229657
electricity	-0.29588784	-0.117298602	-0.007906939	-0.512692772	1.000000000	-0.2403458222	0.143731976	0.528477996	0.421619480	0.205287300	0.0629655760
electricity_ESCOM	0.01833853	0.008759207	0.036318754	0.027241770	-0.240345822	1.0000000000	-0.010896196	-0.023097193	-0.049732025	-0.040403304	-0.0007738747
MTLtelephone	-0.12218527	-0.027982557	0.002756966	-0.071386480	0.143731976	-0.0108961960	1.0000000	0.106836397	0.098066153	0.033471643	0.0192933304
source_drinkingWater	-0.24404832	-0.096144964	0.005288152	-0.494255776	0.528477996	-0.0230971930	0.106836397	1.000000000	0.339072709	0.168273112	0.0332015980

toilet_facility	-0.229374800	-0.101204239	0.006577243	-0.294385356	0.421619480	-0.0497320248	0.098066153	0.339072709	1.000000000	0.243428172	0.0851068874
disposal_facility	-0.109166333	-0.044588504	-0.002301272	-0.176811547	0.205287300	-0.0404033037	0.033471643	0.168273112	0.243428172	1.000000000	0.0877142854
use_bednet	-0.031823633	-0.012454360	-0.001816326	-0.043422966	0.062965576	-0.0007738747	0.019293330	0.033201598	0.085106887	0.087714285	1.0000000000

The above matrix represents the correlation between different variables in our dataset, where we can say that -

1. Correlation between area_property and hous_price is very low , indicating that social factors like electricity, source of water, toilet facilities, etc. have a bigger role to play.
2. Some of the categoricals columns like Electricity have negative correlation with housing price that is because we consider our binary category as (1 : Yes 2 : No)
3. The above explaination can beextended for other categorical columns such as toilet facility, disposal facility, etc.

Extra points

- **Dealing with Null Values** - In our model, we have replaced the Null Values in the dataset with the mean/mode of the other entries in that column respectively. We didn't use Zero(0) to fill these Null values as the errors might be increased when such an assumption is taken.

- **Bootstrapping** - We could use bootstrapping to repeatedly draw sample data in order to avoid the problems with any outliers that might be present. Bootstrapping from the dataset can also help us to build and train models on relatively smaller chunks of data as compared to the original methods.

- **K Fold Cross Validation** - If we were to implement Bootstrapping technique to build and train models on the dataset, we could also use K-Fold Cross validation method in order to get an average value for the Root Mean Square Errors across various algorithms implemented. This reduces dependency on the subset of dataset sampled in Bootstrapping.

- **KNN** - By using KNN for imputation, we can create a more complete dataset that can be used for further analysis. KNN imputation is especially useful when there are only a small number of null values in the dataset, as it can be computationally expensive for large datasets.