Abstract

This report outlines the creation and evaluation of machine learning models designed to forecast housing prices, leveraging a dataset of 12 property characteristics. The suite of models deployed includes a Multivariate linear regression, Ridge Regression, Random Forest, and Radial Support Vector Machines (SVM). Utilizing a dataset comprising 545 price observations from various U.S. states, sourced from Kaggle, the models were assessed for precision using the Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Error (ME), and Root Mean Squared Error (RMSE) on test data. The findings revealed that Random Forest and Radial SVM emerged as the most accurate models for price prediction, due to their proficiency in handling non-linear data regressions.

1. Relevant background and description of the task

Housing price predictive models are important for investment decisions and macroeconomic policy formulation. They enable investors to identify undervalued properties, anticipate market trends, and maximize returns [1]. Government and regulatory bodies rely on these forecasts to craft policies that ensure market stability and address affordability concerns, directly impacting economic well-being and societal equity [2]. Accurate housing price forecasts empower consumers to make informed decisions regarding buying or selling properties [3]. As such, the task of this analysis is to develop machine learning models that can accurately predict housing prices based on property features.

2. Description of the data set

The dataset, sourced from Kaggle, details the prices of 545 houses across various USA states, alongside 12 attributes per property. Attributes include 'price' (sale price), 'area' (property size), 'bedrooms', 'bathrooms', 'stories' (floor count), and 'parking' (parking space count). It also features binary variables ('yes' or 'no') for 'mainroad' (proximity to a main road), 'guestroom', 'basement', 'hotwaterheating', and 'airconditioning', all converted to numerical data for analysis. 'Yes' or 'no' values were encoded as ‘1’ and ‘0’, respectively, while 'furnishingstatus' was categorized into ‘3’ (furnished), ‘2’ (semi-furnished), and ‘1’ (unfurnished).

3. Methodology and Analysis

The dataset for this analysis was split between train and test data following a 70/30 split to test the generated model’s ability to generalise on unseen data. The training data was used for training the Multivariate Linear Regression model (MLR), a Ridge Regression model, a Random Forest model (RF) and a Radial Support Vector Machine model (Radial SVM).

a. Multivariate Linear Regression model (MLR)

Harvey Collier Test was undertaken to measure the linearity between ‘price’ and the other independent variables, which resulted in a high p-value (0.1155). This shows that the null hypothesis on the existence of a linear relationship cannot be rejected. As such, a multivariate linear regression model was selected. The model variables were selected using the best subset regression results to identify the best subset of variables to be used in the model. All 3 best subset methods (i.e., Exhaustive, Forward and Backward selection methods) indicated the use of 8 variables (see Appendix 1). A multivariate linear model of the selected 8 variables (referred to as the subset model) was created and compared with the multivariate linear model including all 12 dependent variables (referred to as the all-variable model). Analysis shows that the all-variable model outperformed the subset model with a higher Adjusted R-square value (70.83% compared to 69.1%) and higher accuracy in generalisation to unseen data (i.e., the all-variable model outperformed the subset models with lower MAE, ME, RSME and MSE values).

Further analysis of the data on why 4 variables were excluded in the subset model showed that the model had a multicollinearity and heteroscedasticity problem. Multicollinearity was identified through correlation analysis which showed a high correlation between independent variables (see Appendix 2). Heteroscedasticity was identified through a residual plot (See Appendix 3) and further investigated through a Breusch-Pagan test, which showed a low p-value (1.746e-07) indicating that the null hypothesis of homoscedasticity can be rejected. A ridge regression model was selected to address the multicollinearity. A ridge regression model was trained through cross-validation to determine the optimal 'lambda' hyperparameter between 10^8 to 10^{-4}. The results of the model outperformed the all-variable MLR. To address Heteroscedasticity, a new dependent variable was created through the logarithmic transformation of the 'price' variable and the ridge regression model was retrained. The resulting ridge regression outperformed the initial ridge regression model as the best-performing linear regression model. In the assessment for generalisation error, the predicted values were transformed back into regular ranges through exponentiation to negate the log transformation. This was done to ensure comparability against other models.

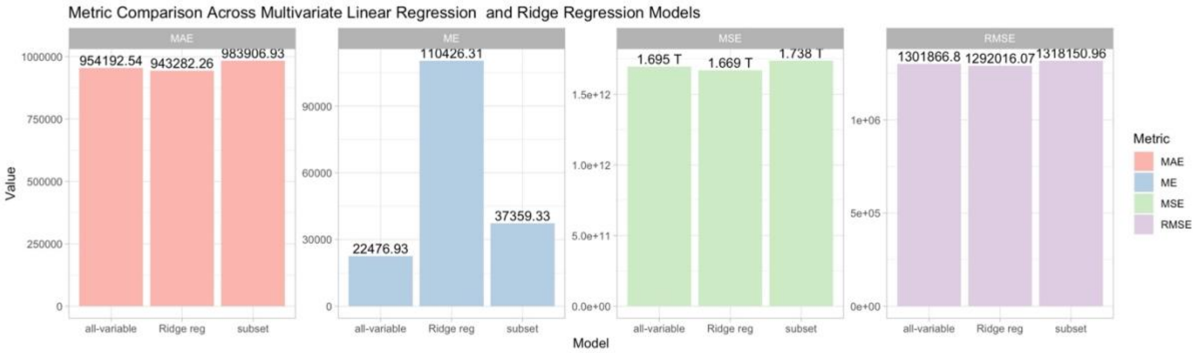


Figure 1: MAE, ME, MSE, and RSME comparison for the Multivariate Linear Regression and the Ridge Regression models. It should be noted that the MSE values are stated in Trillions

b. Radial Support Vector Machine model (Radial SVM)

A Support Vector Machine with a radial base kernel was selected in the analysis to explore the non-linear relationship between house features and ‘price’. The 'cost' and 'gamma' hyperparameters for the Radial SVM were optimised through a grid search algorithm to train multiple SVM models on the training dataset. The 'cost' parameter was set to a range from 2^2 to 2^{16} while the 'gamma' parameter was set to range from 2^{-8} to 2^8. These ranges were selected to explore a wide variety of model behaviours from more rigid to more flexible boundaries between price categories. After exploring a wide parameter range for flexibility in the model, the optimal settings were applied to the final SVM, yielding lower RMSE, MSE, ME and MAE in comparison to the MLR model. The high performance of the Radial SVM indicates a non-linear relationship in the data. However, directly measuring feature importance from Radial SVM is computationally expensive. Therefore, a Random Forest was used to further analyse these non-linear relationships and provide feature importance scores.

c. Random Forest model (RF)

The focus of training the model was on accuracy thus the number of trees estimators were not limited. To select the best Random Forest model, multiple random forest models were created to means to tune the “mtry” hyperparameter, which indicates the number of features that the algorithm will randomly choose the best split from those features. The best model selected was “mtry” = 9, as it had the lowest MSE, ME, RMSE and MAE (see Figure 2).

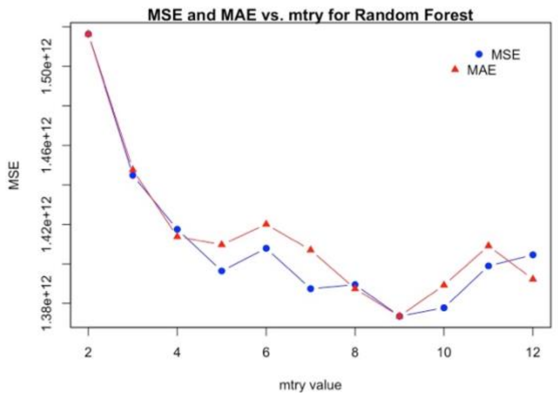


Figure 2: MAE vs MSE for numerous values of "mtry" hyperparameter plotted on two different axes

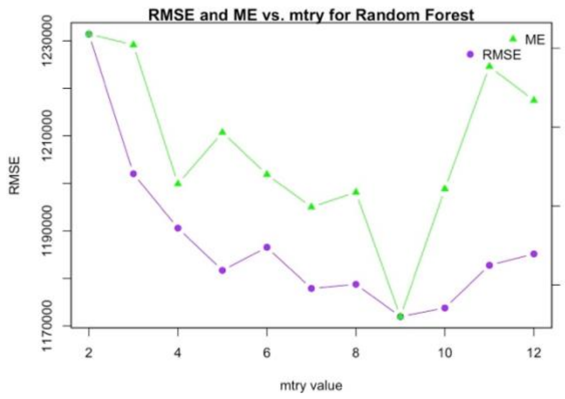


Figure 3: RMSE vs ME for numerous values of "mtry" hyperparameter plotted on two different axes

Upon further analysis of the contribution of each variable to the Random Forest model, it was highlighted ‘hotwaterheating’ feature had the lowest contribution to the model, followed by ‘parking’, ‘guestroom’, and ‘mainroad’ (see Appendix 5). This contradicts the findings from the MLR analysis which shows that showed ‘bedrooms’ feature has the lowest importance in price prediction, indicating that there is a significant non-linear relationship between bedrooms and ‘price’ that is better captured through the random forest model. Additionally, the ‘guestroom’ feature has low importance as a variable in the RF model and similarly in the MLR model. This shows that ‘guestroom’ has low input in the price prediction due to its low linear and non-linear relationship to the target feature (i.e., ‘price’). In dealing with large datasets, features with marginal contributions to predictive accuracies, such as 'hotwaterheating' and 'guestroom', are often omitted to streamline the model and reduce computational demand. These features, while potentially informative, can disproportionately increase the complexity of the model relative to the value of insight they provide. However, in the context of this analysis, the dataset is relatively smaller, and the computational cost incurred by including these features is minimal. Given this, it is justifiable to retain them, as they enhance the prediction accuracy. This decision allows us to balance the desire for a more refined model against the practical considerations of computational efficiency.

4. Results Evaluation

The Random Forest (RF) model consistently outperforms the other three models across all reported metrics (MAE, MSE, ME, and RMSE). Its lower values indicate better prediction accuracy and reliability. While the data exhibited linear tendencies, the presence of non-linear patterns significantly influenced the target variable. This is reflected by the superior performance of RF and Radial SVM models, both of which effectively excel in managing non-linear relationships. The best Ridge Regression model with dependent variable transformation had the second-highest error rates. This suggests that including all variables without accounting for non-linear relationships does not yield the most accurate predictions. Further evaluation of mean error values reveals that the Radial SVM and Ridge Regression tend to underestimate house prices, indicated by a higher Mean Error, whereas the RF model displays a more balanced error distribution, suggesting a more reliable performance across various price levels. Thus, the most effective machine learning model for predicting house prices is one capable of capturing the intricate non-linear relationships within the data, with RF being the preferred model in this scenario.

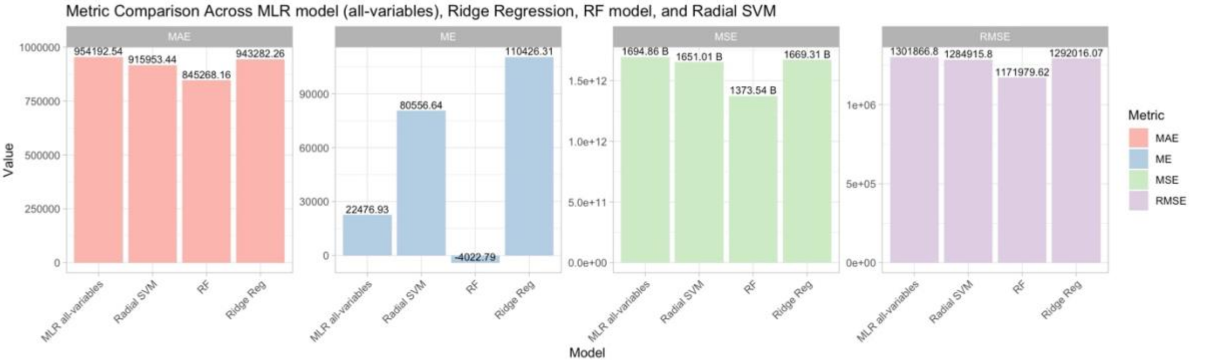


Figure 4: MAE, ME, MSE, and RSME comparison for the best models across MLR, RF, and Radial SVM models. It should be noted that the MSE values are stated in Billions for readability

Appendix

Appendix 1: Best subset regression summary

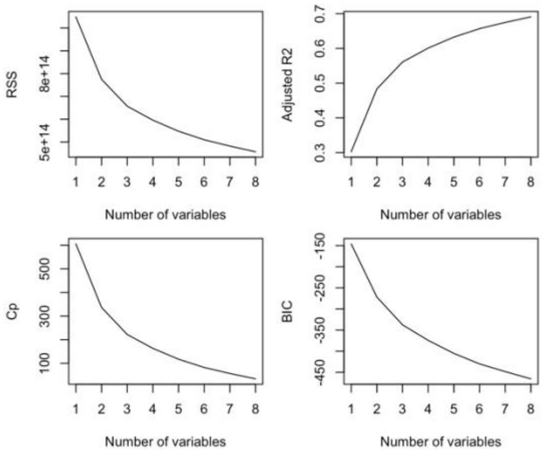


Figure 5: Best subset regression results using Exhaustive method

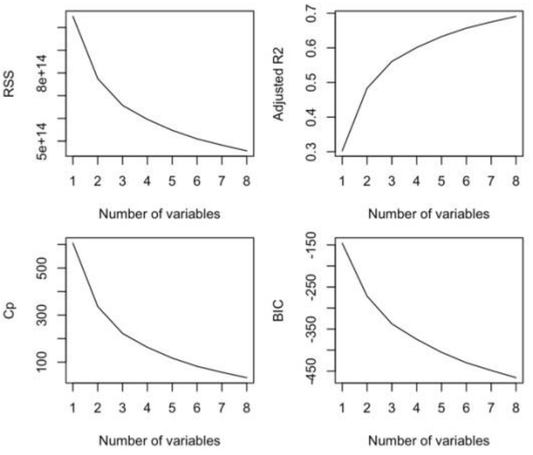


Figure 6:Best subset regression results using Forward method

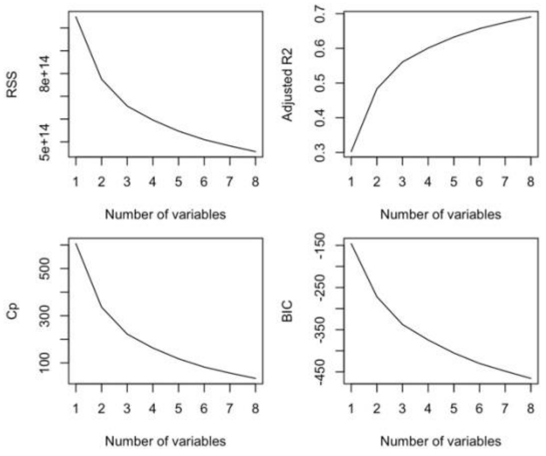
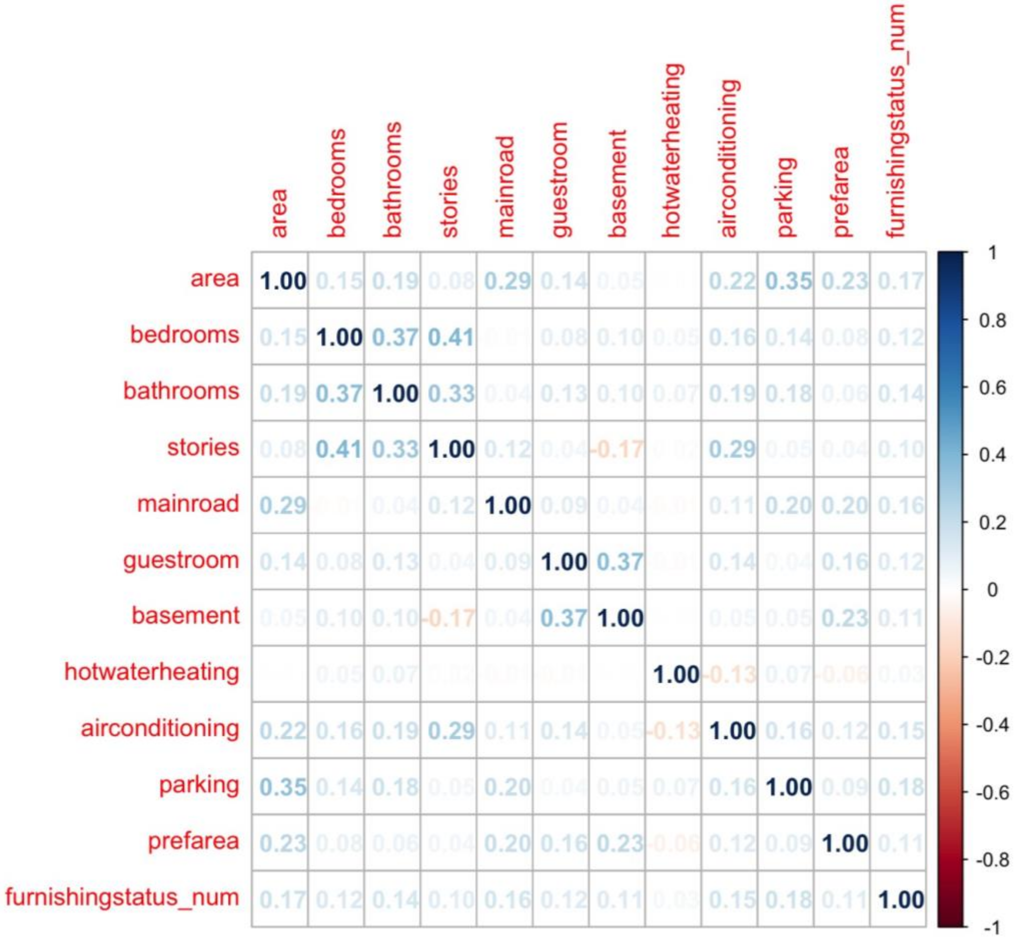


Figure 7: Best subset regression results using Backward method

Appendix 2: Correlation Analysis

Multicollinearity exists in the dataset evident in the high independent variable correlation with one another. These include:

- Area and Mainroad at ~29%
- Area and Parking at ~35%
- Bedrooms and Bathroom at ~37%
- Bedrooms and Stories at ~41%
- Bathrooms and Stories at ~33%
- Stories and Airconditioning at ~29%
- Guestroom and Basement at ~37%



	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus_num
area	1.00	0.15	0.19	0.08	0.29	0.14	0.05		0.22	0.35	0.23	0.17
bedrooms	0.15	1.00	0.37	0.41	-0.01	0.08	0.10	0.05	0.16	0.14	0.08	0.12
bathrooms	0.19	0.37	1.00	0.33	0.04	0.13	0.10	0.07	0.19	0.18	0.06	0.14
stories	0.08	0.41	0.33	1.00	0.12	0.04	-0.17	0.02	0.29	0.05	0.04	0.10
mainroad	0.29	-0.01	0.04	0.12	1.00	0.09	0.04	-0.01	0.11	0.20	0.20	0.16
guestroom	0.14	0.08	0.13	0.04	0.09	1.00	0.37	-0.01	0.14	0.04	0.16	0.12
basement	0.05	0.10	0.10	-0.17	0.04	0.37	1.00		0.05	0.05	0.23	0.11
hotwaterheating		0.05	0.07	0.02	-0.01	-0.01		1.00	-0.13	0.07	-0.06	0.03
airconditioning	0.22	0.16	0.19	0.29	0.11	0.14	0.05	-0.13	1.00	0.16	0.12	0.15
parking	0.35	0.14	0.18	0.05	0.20	0.04	0.05	0.07	0.16	1.00	0.09	0.18
prefarea	0.23	0.08	0.06	0.04	0.20	0.16	0.23	-0.06	0.12	0.09	1.00	0.11
furnishingstatus_num	0.17	0.12	0.14	0.10	0.16	0.12	0.11	0.03	0.15	0.18	0.11	1.00

Figure 8: Correlation among the independent variables

Appendix 3: Homoscedasticity Analysis through analysing the residuals

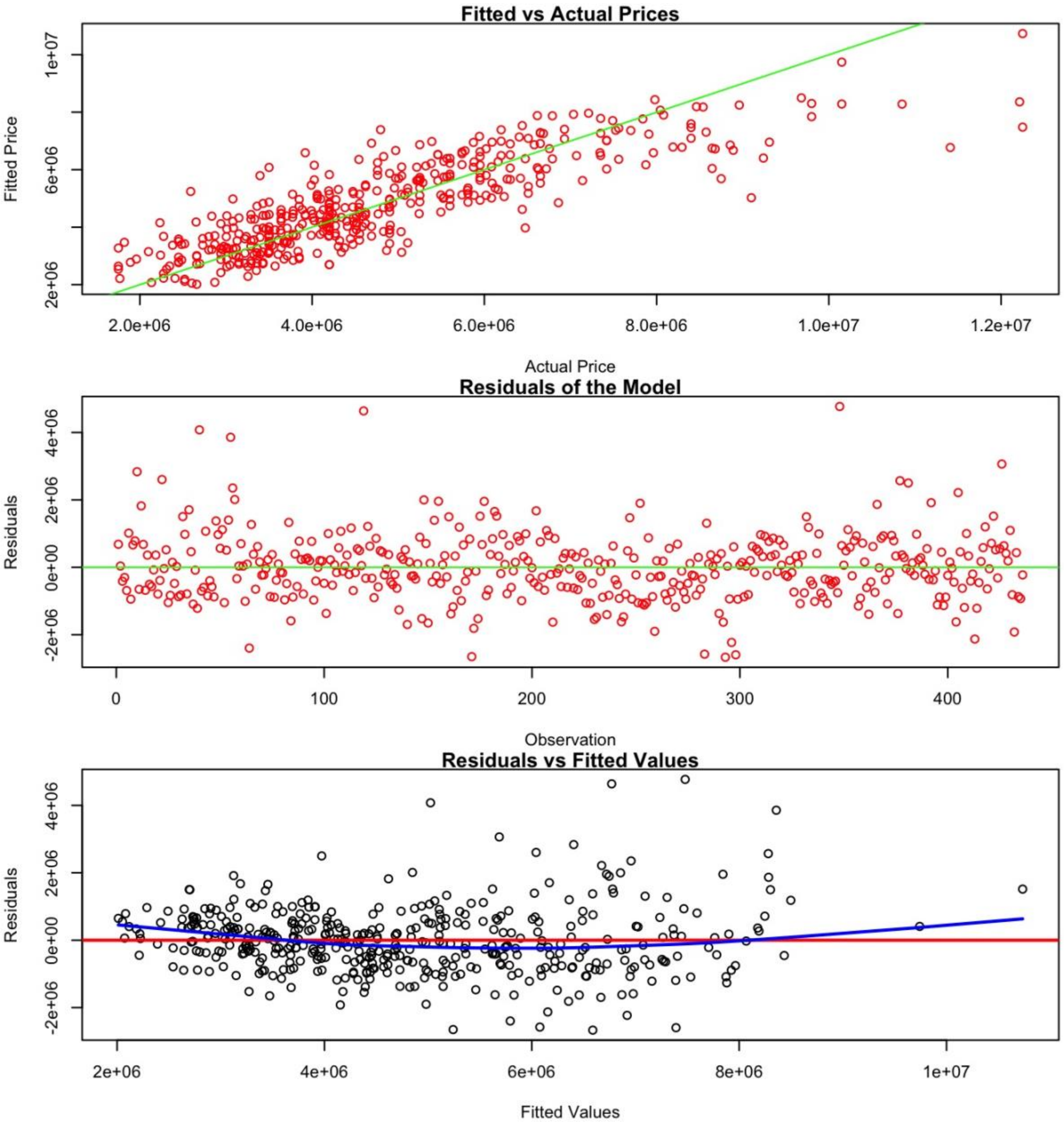


Figure 9: A residual analysis using Fitted vs Actual price scatter plot, Residuals scatter plot, and Residual vs Fitted Value scatter plot

Appendix 4: Cross-validation results for Ridge regression

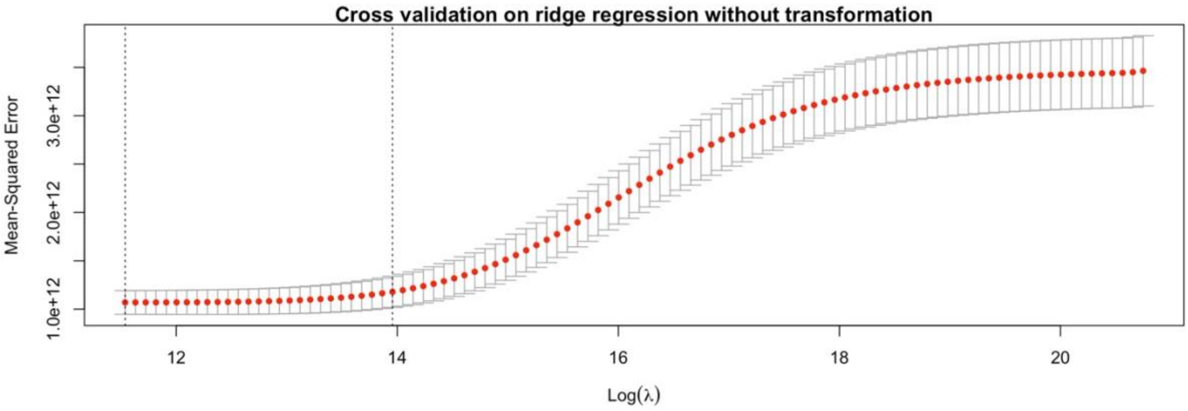


Figure 10: Cross-validation results from Ridge Regression without transformation of dependent variable (i.e., 'price')

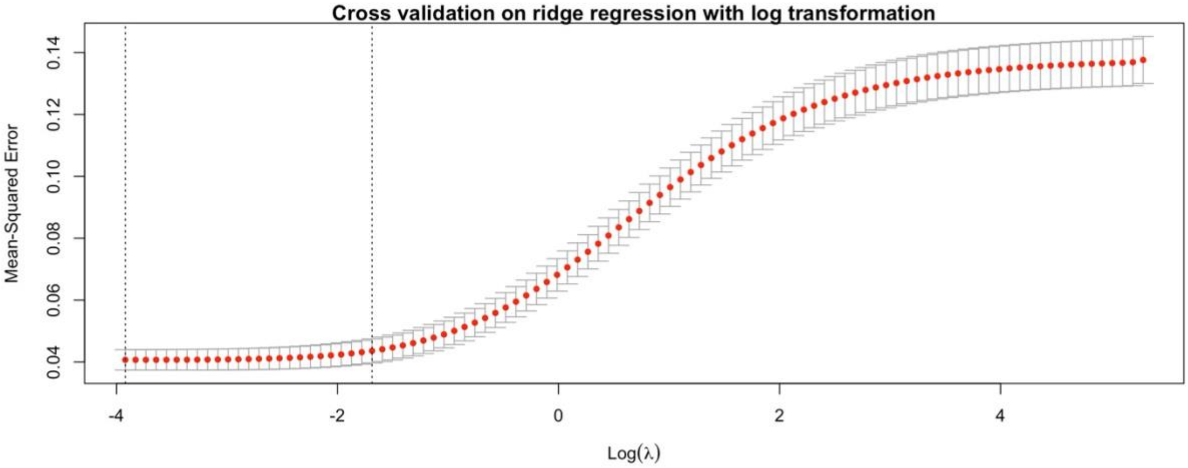


Figure 11: Cross-validation results from Ridge Regression with logarithmic transformation of dependent variable (i.e.,log 'price')

Appendix 5: Feature importance Analysis of the Random Forest Model

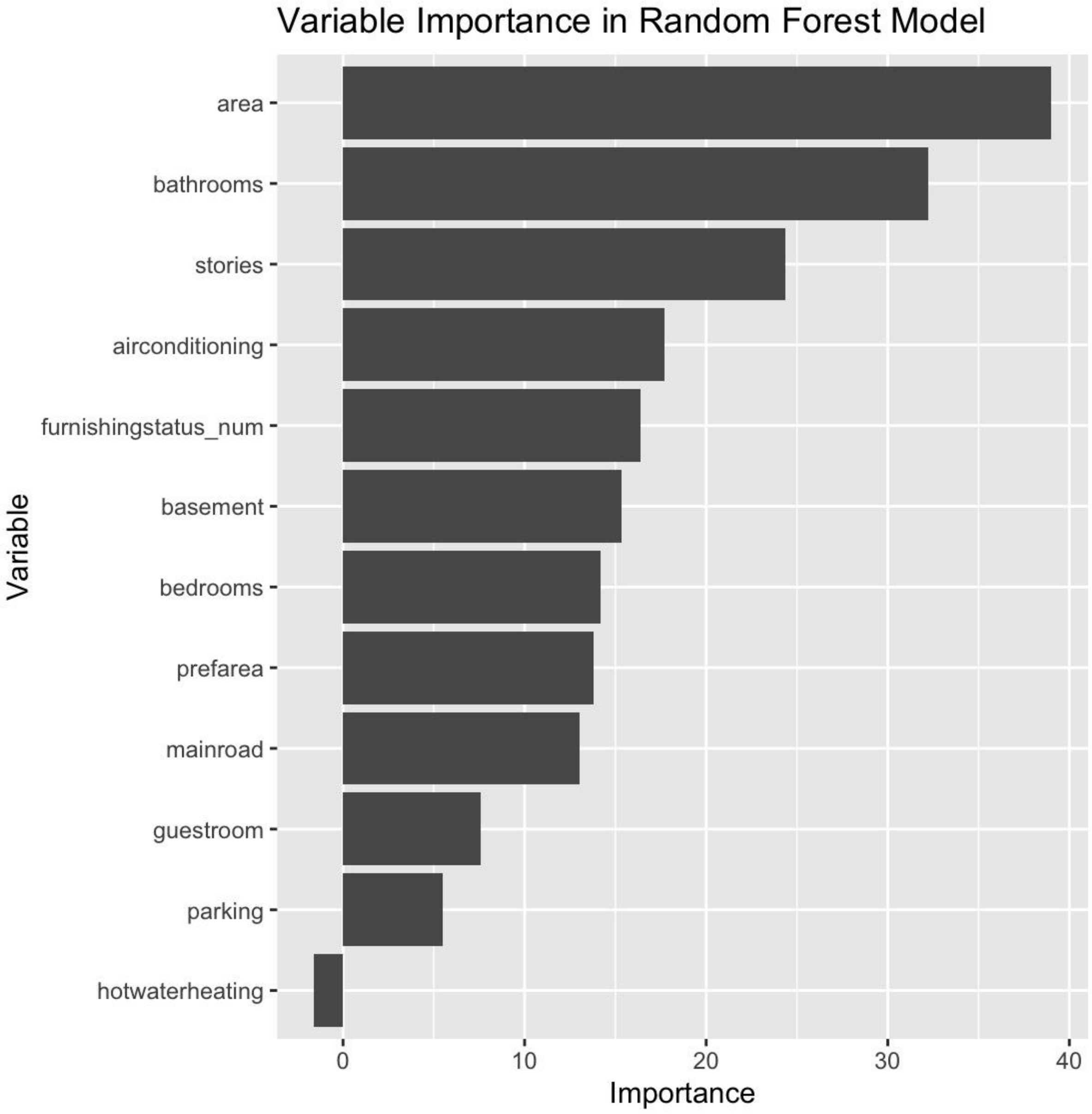


Figure 12: Variable importance of the Random Forest model with 'mtry' = 9

References

1. Goodhart, C. A. E., & Hofmann, B. (2008). House prices, money, credit, and the macroeconomy. Oxford Review of Economic Policy, 24(1), 180-205.
2. Case, K. E., & Shiller, R. J. (2003). Is there a bubble in the housing market? Brookings Papers on Economic Activity, 2003(2), 299-342.
3. Shiller, R. J. (2007). Understanding recent trends in house prices and homeownership. NBER Working Paper Series.