World Academy of Science, Engineering and Technology
International Journal of Business and Economics Engineering
Vol:11, No:4, 2017

Feature Selection and Predictive Modeling of Housing Data Using Random Forest

Bharatendra Rai

***Abstract*—**Predictive data analysis and modeling involving machine learning techniques become challenging in presence of too many explanatory variables or features. Presence of too many features in machine learning is known to not only cause algorithms to slow down, but they can also lead to decrease in model prediction accuracy. This study involves housing dataset with 79 quantitative and qualitative features that describe various aspects people consider while buying a new house. Boruta algorithm that supports feature selection using a wrapper approach build around random forest is used in this study. This feature selection process leads to 49 confirmed features which are then used for developing predictive random forest models. The study also explores five different data partitioning ratios and their impact on model accuracy are captured using coefficient of determination (r-square) and root mean square error (rsme).

***Keywords*—**Housing data, feature selection, random forest, Boruta algorithm, root mean square error.

I. Introduction

PREDICTIVE modeling in presence of a large number of exploratory variables requires use of methods that support feature selection. Random forest algorithm is a popular machine learning method that automatically calculates variable importance measure as a by-product and has been successfully used by various researchers [1], [2]. Variable importance measures that include mean decrease accuracy (MDA) and mean decrease Gini (MDG) provided by random forest have also been studied for stability. Studies involving simulations indicate that ranks based on MDA are unstable to small perturbations of the dataset whereas ranks based on MDG provide more stable results [3]. At the same time in situations where there are strong within-predictor correlations, MDA rankings are found to be more stable that MDG [4]. It is also known that having too many features can not only slow down algorithms, but many machine learning algorithms also exhibit a decrease in accuracy in such situations [5].

Boruta algorithm that supports feature selection uses a wrapper approach build around random forest methodology [6]. An output of Boruta algorithm provides classification of explanatory variables or features into three categories, viz., important, tentative, and unimportant variables or features. It also allows a rough fix for tentative variables which can be used to fill missing decisions regarding importance or unimportance by simple comparison of the median attribute Z score with the median Z score of the most important shadow attribute. Many researchers have successfully applied this algorithm that provides several advantages in feature selection to support predictive modeling [7], [8].

This paper provides an application of Boruta algorithm for feature selection and then uses random forest algorithm for predictive modeling of housing data involving 79 explanatory features. These features describe various aspects people consider while buying a new house. The main objective of this study is to develop a predictive model for the sale price of the house based on appropriate features.

B. Rai is with the Charlton College of Business, University of Massachusetts-Dartmouth, North Dartmouth, MA 02747 USA (phone: 508-910-6434; e-mail: brai@umassd.edu).

II. Feature Selection from Exploratory Variables in the Housing Data

A. Data for the Study

The dataset used for this study consists of data on 1460 houses with 79 exploratory variables and property's sale price in dollars as target variable based on the location of Ames city in Iowa State, USA. This dataset was made available through a competition on kaggle.com. There are 43 qualitative, 31 quantitative and 4 date related variables out of 79 exploratory variables. A heatmap based on correlation coefficients of quantitative variables is shown in Fig. 1.

B. Missing Data

There are 18 variables with missing data that range from a low of 8 to as high as 1406. The missing values for quantitative variables are replaced using average value of that variable and missing values for qualitative variables are replaced by zero representing another level for that variable.

C. Features Selection

Boruta package available in R software is used for feature selection in this study. It uses a wrapper algorithm and can work with any classification methodology that yields variable importance measure (VIM) as an output and by default uses random forest. This analysis performed 100 iterations in a total of about 7.13 minutes. The results yielded 49 attributes confirmed as important, 19 attributes confirmed as unimportant, and 11 tentative attributes as shown in Fig. 2. In Fig. 2, boxplots that are green in color represent features classified as important, yellow boxplots represent tentative features and red boxplots represent unimportant features. Top three features based on the analysis are above ground living area in square feet (GrLivArea), overall material & finish quality (OverallQual), and second floor square feet (X2ndFlrSF) based on maximum importance values of 24.34, 19.74, and 17.91 respectively. The output from the analysis also provides mean, median, maximal and minimal

importance, number of hits normalized to number of importance source runs performed and the decision about feature importance. Fig. 3 provides a plot of number of hits normalized to number of importance source runs performed versus mean importance for the feature categories.

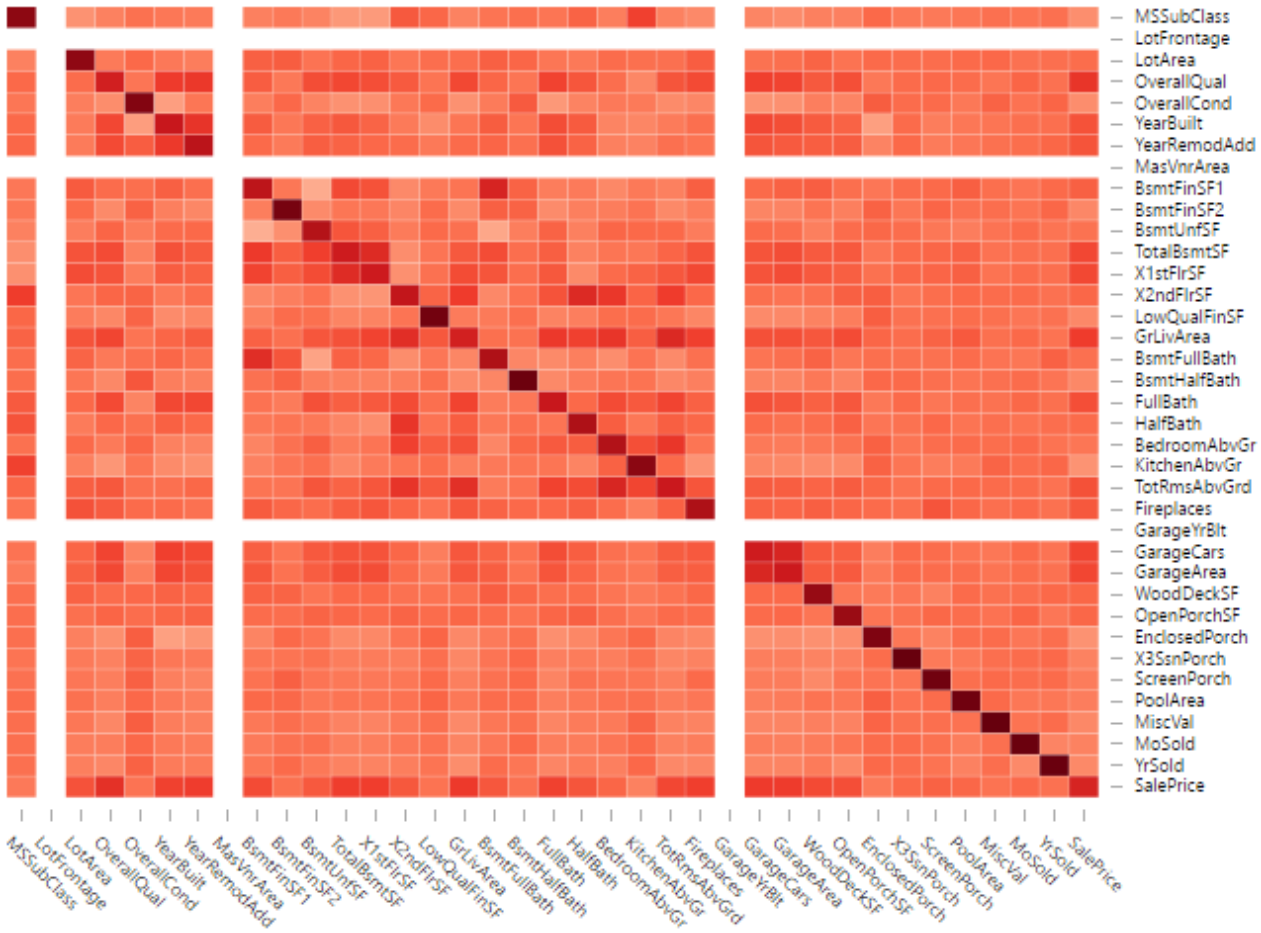

Fig. 1 Correlation coefficient heatmap of quantitative variables

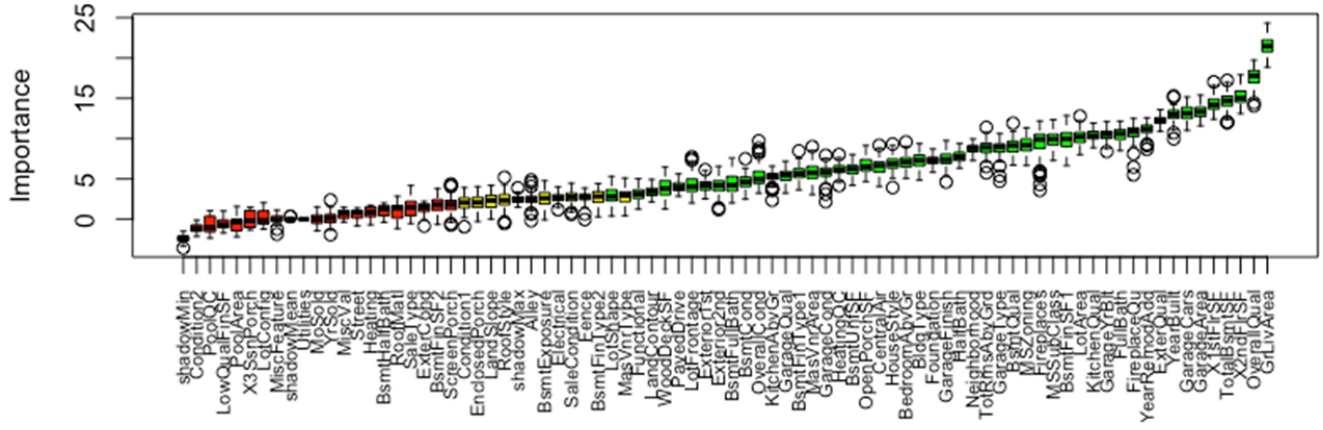

Fig. 2 Important, tentative, and unimportant features based on Boruta analysis

Using tentative rough fix, a final classification of 79 features into 57 as important and 22 as unimportant is arrived at. Feature groupings based on Boruta analysis are summarized in Table I.

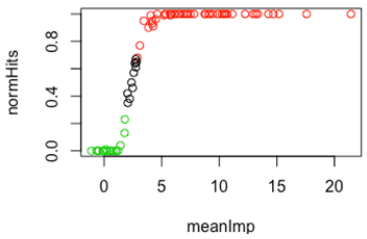

Fig. 3 Number of hits normalized to number of importance source runs performed versus mean importance

TABLE I
THREE DIFFERENT GROUPING OF FEATURES

Feature Groupings	Features Included
79 features	All 79 features included
49 features originally confirmed by Boruta analysis	MSSubClass + MSZoning + LotFrontage + LotArea + LotShape + LandContour + Neighborhood + BldgType + HouseStyle + OverallQual + OverallCond + YearBuilt + YearRemodAdd + Exterior1st + Exterior2nd + MasVnrArea + ExterQual + Foundation + BsmtQual + BsmtCond + BsmtFinType1 + BsmtFinSF1 + BsmtUnfSF + TotalBsmtSF + HeatingQC + CentralAir + X1stFlrSF + X2ndFlrSF + GrLivArea + BsmtFullBath + FullBath + HalfBath + BedroomAbvGr + KitchenAbvGr + KitchenQual + TotRmsAbvGrd + Functional + Fireplaces + FireplaceQu + GarageType + GarageYrBlt + GarageFinish + GarageCars + GarageArea + GarageQual + GarageCond + PavedDrive + WoodDeckSF + OpenPorchSF
57 features based on tentative rough fix	MSSubClass + MSZoning + LotFrontage + LotArea + Alley + LotShape + LandContour + Neighborhood + BldgType + HouseStyle + OverallQual + OverallCond + YearBuilt + YearRemodAdd + RoofStyle + Exterior1st + Exterior2nd + MasVnrType + MasVnrArea + ExterQual + Foundation + BsmtQual + BsmtCond + BsmtExposure + BsmtFinType1 + BsmtFinSF1 + BsmtFinType2 + BsmtUnfSF + TotalBsmtSF + HeatingQC + CentralAir + X1stFlrSF + X2ndFlrSF + GrLivArea + BsmtFullBath + FullBath + HalfBath + BedroomAbvGr + KitchenAbvGr + KitchenQual + TotRmsAbvGrd + Functional + Fireplaces + FireplaceQu + GarageType + GarageYrBlt + GarageFinish + GarageCars + GarageArea + GarageQual + GarageCond + PavedDrive + WoodDeckSF + OpenPorchSF + EnclosedPorch + Fence + SaleCondition

III. RANDOM FOREST PREDICTION MODELS

Random forests are extension of the idea of decision trees [9], [10]. Unlike a single tree that is constructed in decision tree, multiple decision trees are constructed leading to a random forest. The output from all trees is combined to obtain a better model than what could be obtained from a single tree. Random forest models can be used for developing classification models when the response variable is a factor and can also be used for developing a prediction model when response variable is continuous as in this study. The model is developed using *randomForest* package available from R software. Random forest has two free parameters viz., number of trees (ntree) and number of variables randomly sampled as candidates at each split (mtry). The default value for ntree is 500 trees in the random forest and default value for mtry is about p/3 for regression where p is the number of features. Coefficient of determination (r-square) and root mean square error (RMSE) are used for assessing the performance of the prediction model. Fig. 4 shows error rate of a random forest model based on all 79 features with house sale price as the dependent variable.

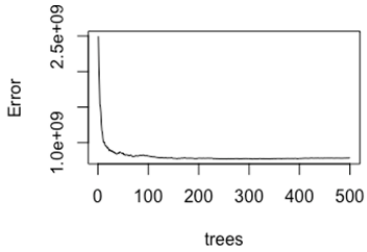

Fig. 4 Error rate of a random forest model with all 79 features

It can be observed from Fig. 4 that the error rate becomes flat after about 150 trees. This indicates that increasing the number of trees beyond the default value of 500, is unlikely to have a significant impact on the model accuracy. Therefore, for this study the default value for number of trees is kept constant at 500.

Data partitioning with 50:50, 60:40, 70:30, 80:20, and 90:10 splits into training and testing datasets respectively are used in the study. Random forest models are developed for three different feature groupings involving all 79 features, 49 originally confirmed features and 57 features confirmed with tentative rough fix. A random forest model is built using the training dataset. To enable consistency in comparison of results across various training and testing datasets for each of the three feature groupings, a random seed with set.seed(123) is fixed for each data partitioning split. R-square and RMSE calculated using training dataset and RMSE calculated using testing datasets are used for model assessment. Higher values of r-square and lower values of RMSE are desired. The results obtained for five different ratios of training and testing data splits, and three different groupings of features are summarized in Table II.

TABLE II
PERFORMANCE OF RANDOM FOREST MODELS FOR FIVE DIFFERENT RATIOS OF TRAINING AND TESTING DATA SPLITS AND THREE DIFFERENT GROUPING OF FEATURES

Split	Variables	Train Data R-Sq	Train Data RMSE	Test Data RMSE
50:50	79 - All	85.83	29132.07	31603.16
50:50	49 - Confirmed	87.04	27860.26	31811.52
50:50	57 - Confirmed with rough fix	86.32	28619.47	31633.36
60:40	79 - All	87.91	27398.04	34497.04
60:40	49 - Confirmed	87.98	27319.26	34130.56
60:40	57 - Confirmed with rough fix	87.55	27800.49	34644.65
70:30	79 - All	86.93	29668.28	27015.77
70:30	49 - Confirmed	87.21	29353.45	26554.12
70:30	57 - Confirmed with rough fix	87.03	29554.35	26395.83
80:20	79 - All	87.31	28522.43	29162.15
80:20	49 - Confirmed	87.57	28226.59	28648.07
80:20	57 - Confirmed with rough fix	87.41	28402.20	29154.09
90:10	79 – All	87.64	28026.23	27663.77
90:10	49 - Confirmed	87.54	28144.27	25982.38
90:10	57 - Confirmed with rough fix	87.97	27652.51	27000.87

Table II shows that r-square values based on training

dataset are consistently higher when 49 features originally confirmed by the Boruta analysis are used except when training and testing splits are 90:10. The highest r-square value of 87.98% is obtained with 60:40 split for 49 feature grouping. Similarly, RMSE based on training dataset are consistently lower when 49 features originally confirmed by the Boruta analysis are used except when the splits are 90:10. The lowest RMSE value of 27319.26 is obtained with 60:40 split for 49 feature grouping. For 60:40 split, RMSE value is also lower for testing data when 49 features originally confirmed by Boruta analysis are used. These results indicate that use of unimportant and tentative variables in the random forest prediction model for house sale price do not help to improve model accuracy. The results also suggest that data partitioning ratio used for model development and assessment may also influence model accuracy. Although for the dataset used in this study 60:40 split provides better model accuracy, for a different dataset some other ratio may result in a better accuracy levels.

IV. Discussion of the Results

Random forest models developed in the previous section suggested 49 confirmed features based on Boruta analysis and 60:40 split provides improved model accuracy. Note that the number of trees in this random forest model is 500. Fig. 5 shows a histogram of tree size or number nodes in each of the 500 trees in the random forest model.

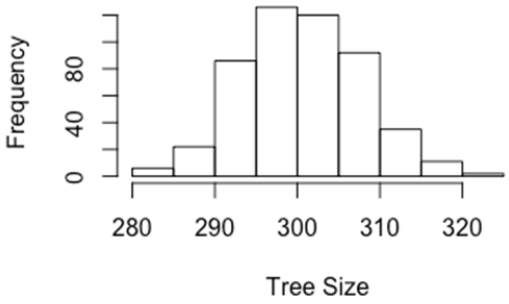

Fig. 5 Histogram of tree size or number of nodes per tree in the random forest model based on 49 confirmed features

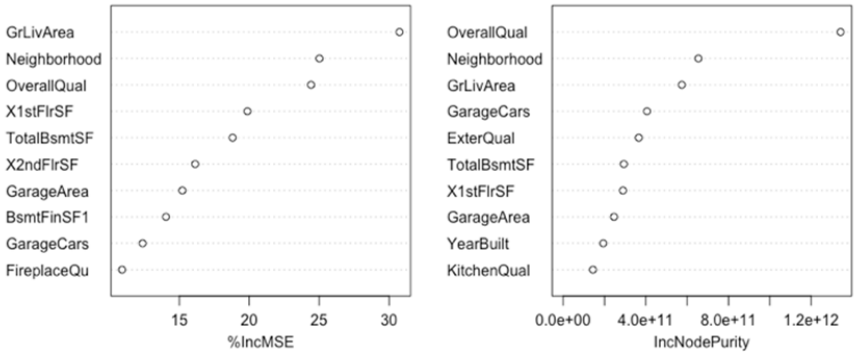

Fig. 6 Top ten variable importance plot based on the random forest model from 49 confirmed features

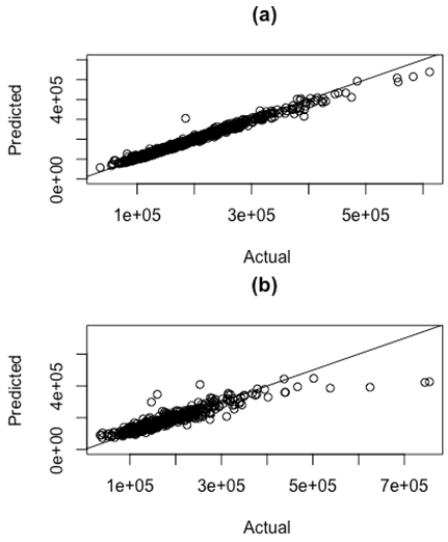

Fig. 7 Actual versus predicted sales price based on the random forest model from 49 confirmed features, (a) training data, and (b) testing data

Fig. 5 shows on an average there are about 300 nodes in a tree for the random forest model. The number nodes vary from about 280 to 325 per tree. The shape of the histogram is approximately symmetrical.

Fig. 6 provides the variable importance plot using randomForest package in R based on the random forest model from 49 confirmed features.

The variable importance plot in Fig. 6 shows what impact each feature has if removed from the model. The importance is captured using percentage increase in mean square error (MSE) and increase in node purity. Removing GrLivArea from the random forest model has the highest impact on percentage increase in MSE. Similarly, dropping OverallQual has maximum impact on node purity. Note that this list of top ten features is from among 49 confirmed features based on Boruta analysis. In addition, these two features were also in the top two list in importance for Boruta analysis.

The performance of the random forest model is further assessed using training and testing dataset respectively as shown in Figs. 7 (a) and (b).

Fig. 7 (a) shows a decent fit between actual and predicted house sales price based on the training dataset. For sales price over $500K, there seems to be under-estimation in the sale of

house prices. Such under-estimation towards higher house prices is seen even more in the testing dataset. This analysis and results also suggests need for exploring other machine learning algorithms such as neural networks or support vector machines to further improve the prediction accuracy [11].

V. Conclusions

In this study, feature selection approach involving Boruta algorithm is illustrated using housing data. Random forest models using three feature groupings involving all 79 features, 49 features confirmed by Boruta analysis, and 57 features using tentative rough fix are developed. Results obtained indicate better model accuracy in terms of r-square and RMSE for feature grouping with 49 confirmed features based on Boruta analysis. Although data partitioning with 60:40 split performed comparatively better than other four split ratios used, results also suggest scope for further improving the model accuracy by exploring other machine learning approaches. This is especially true for house prices that are above $500K.

References

[1] M. J. Hallett, J. J. Fan, X. G. Su, R. A. Levine, and M. E. Nunn, "Random forest and variable importance rankings for correlated survival data, with applications to tooth loss," *Statistical Modelling*, Vol.14(6), pp.523-547, 2014.

[2] A. L. Boulesteix, A. Bender, B. J. Lorenzo, and C. Strobl, "Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations," *Briefings in Bioinformatics*, Vol. 13(3), pp.292-304, 2012.

[3] M. L. Calle, and V. Urrea, "Letter to the Editor: Stability of Random Forest importance measures," *Briefings in Bioinformatics*, Vol. 12(1), pp.86-89, 2011.

[4] K. Nicodemus, "Letter to the Editor: On the stability and ranking of predictors from random forest variable importance measures, "*Briefings in Bioinformatics*, Vol.12(4), pp.369-373, 2011.

[5] R. Kohavi, and G. H. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence*, 97, 273–324, 1997.

[6] M. B. Kursa, and W. R. Rudnicki, "Feature Selection with the Baruta Package," *Journal of Statistical Software*, Vol. 36, Issue 11, 1-13, 2010.

[7] M. B. Kursa, "Robustness of Random Forest-based gene selection methods," *BMC Bioinformatics*, Vol.15, pp. 1-8, 2014.

[8] Z. Yang, M. Jin, Z. Zhang, J. Lu, and K. Hao, "Classification Based on Feature Extraction for Hepatocellular Carcinoma Diagnosis Using High-throughput DNA Methylation Sequencing Data, "Procedia Computer Science, Vol.107, pp.412-417, 2017.

[9] D. T. Larose, and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, New Jersey: John Wiley & Sons, 2014.

[10] G. Shmueli, N. R. Patel, and P. C. Bruce, *Data Mining for Business Intelligence: Concepts, Techniques, and Applications*. Hoboken, New Jersey: John Wiley & Sons, 2010.

[11] B. K. Rai, "Classification, Feature Selection and Prediction with Neural-Network Taguchi System," *International Journal of Industrial and Systems Engineering*, Vol. 4, No. 6, 645-664, 2009.

Bharatendra Rai is PhD in industrial engineering from Wayne State University, Detroit in 2004. He has master of technology degree in quality, reliability, and operations research from Indian Statistical Institute, India in 1993, and another master's degree in statistics from Meerut University, India in 1991.

He is currently chair and associate professor in the department of decision and information sciences at Charlton College of Business, University of Massachusetts-Dartmouth. He has earlier worked as a quality and reliability engineer at Ford Motor Company in Detroit. He has co-authored a book titled 'Reliability Analysis and prediction from Warranty Data: Issues, Strategies, and Methods' (New York, NY: CRC Press Taylor & Francis Company, 2009). He is author of over 30 journal and conference articles including two publications in IEEE Transactions on Reliability.

Dr. Rai is a certified six-sigma black belt from American Society for Quality. Dr. Rai is also awarded certificates as ISO 9000 and ISO 14000 lead assessor by British Standards Institute and Marsden Environmental International respectively.