Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 319–327
http://www.ssca.org.in/journal

A Hybrid Regression Model for Cashew Nuts Price Prediction

Satyanarayana[1] **and Ismail B.**[2]
[1] *Department of Statistics, Mangalore University, Mangalagangothri, India*
[2] *Department of Statistics, Yenepoya (Deemed to be University), Mangalore, India*

Received: 24 June 2022; Revised: 14 January 2023; Accepted: 17 April 2023

Abstract

This paper proposes a hybrid regression model based on the regression tree and multiple linear regression model for improving prediction accuracy and to overcome one of the main disadvantages of the Regression Tree. The performance of the proposed model is compared with regression tree, K-nearest neighbor regression, multiple linear regression, and support vector regression through a Monte-Carlo simulation study. The simulation result indicates that the hybrid model outperforms all other regression models irrespective of sample size when the observations are from a normal distribution and uniform distribution. As an application, the proposed hybrid model is used to solve a problem faced by cashew nuts farmers and buyers to decide the most appropriate prices for the cashew nuts. The results from the hybrid model can be used as a guide by the farmers for fetching better prices in the market and by buyers for getting a lot of ascertained quality.

Key words: Cashew nuts price; Hybrid model; Regression tree; Support vector regression.

AMS Subject Classifications: 62J05

1. Introduction

Improving efficiency the prediction accuracy of regression model is still an interesting topic for researchers due to the natural variations in the systems themselves, which may drastically affect the model performance. In a traditional regression model, one must make assumptions about the functional form that connects the response variable with explanatory variable(s) which may not be valid. Most of the non-parametric regression techniques depend on the appropriate kernel or bandwidth selection and do not perform well in the case of high dimensional. CART (Classification and Regression Tree) is the most popular, efficient and widely used method for constructing decision trees introduced by Breiman *et al.* (1984). Shih (1997, 2004) observe that the splitting procedure in Regression Tree (RT) is biased as it searches for all possible splits and suggests that a proper normalization method will overcome this difficulty.

Corresponding Author: Satyanarayana
Email: sathya1301@gmail.com

The main disadvantages of RT are

a. RT assigns the same predicted value, average value, for all the tuples in a branch that satisfies the same corresponding splitting criterion.

b. Sometimes RT over fit the datasets, i.e., model completely fits the train data but fails to generalize for the test data.

To overcome the problem of over fitting, a sequence of values for threshold parameters are considered. According to cross-validation technique, the final threshold value is selected based on minimum prediction error criterion. Alternatively, one can also select the final threshold value by using the 1-standard error rule, which yields a prediction error of one standard deviation larger than the minimum error estimated by the cross-validation method. CART has several advantages over the traditional regression model.

The review paper of Domor *et al.* (2019) highlights the prediction performance of various decision tree algorithms. They also carried out an in-depth review of various methods used to improve the performance of the algorithms. Many papers have appeared in the direction of a hybrid modelling approach to improve prediction accuracy. Bennett *et al.* (1998) proposed a Support Vector Machine (SVM) approach to a decision tree to build a hybrid model. Kumar and Gopal (2010) hybrid SVM model-based decision tree and Chang and Liu's (2012) decision tree as an accelerator for SVM are the noticeable works in this direction. Muhamad Safiih Lola *et al.* (2016) proposed a hybrid model based on Artificial Neural Network and Multiple Linear Regression model (MLR). They showed that hybrid approach could improve the performance of Multiple Linear Regression model. Tanujit Chakraborty (2019) proposed a hybrid regression model based on Regression Tree and support vector regression for boiler water quality prediction. Regression Tree can model the arbitrary decision boundaries and found to be more robust algorithm. It has a built-in variable selection method and also it can handle missing values.

The proposed approach is similar to local linear regression using the bandwidth method. Here instead of computing bandwidth to fit regression line locally, the linear regression model is fit to each branch separately after arranging the observations according to splitting criterion. Since observations in each branch show high intra class similarity, the fitted model is expected to perform better than the linear model because the linear regression line is fitted globally. In the hybrid model, the strength of Regression Tree is used to improve the strength of the Multiple Linear Regression model. The proposed model can be used to select the best subset of regressors and for the prediction task. It has the advantages of significant accuracy and easy interpretability.

This work is motivated by a problem faced by cashew nuts buyers and the sellers to decide the most appropriate price for the cashew nuts. The price of cashew nuts can be decided from several quality measurements on the raw and kernel of the cashew nuts. The quality of cashew nuts brought to the market by the farmers varies considerably from lot to lot. In the case of farmers, if the quality of grown cashew nuts is good but due to lack of proper assessment about their quality, they may sell their whole lot for a lesser price. From the point of buyers, after offering a reasonable price for the raw cashew nuts, if the buyers do not get good quality kernels after de-shelling raw cashew nuts, it leads to massive losses because raw cashew nuts are purchased in a large number of lots. Also, the process of producing kernels ready for marketing involves a large amount of human resources. Therefore, it is essential to develop a model which accesses the quality of the cashew nuts with minimal effort and decides optimal remunerative prices for the lot. The cashew nut plays a vital role in economic activities because the cultivation and marketing of cashew nuts involve a considerable amount of manpower in India. India is the largest producer of cashew nuts in the world. The problems associated with its cultivation, trading and marketing are that the growers do not reap optimal return and traders do not get reasonable profit.

2. Methodology

a. MLR method

Consider the multiple regression model $Y = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$, where Y is an $n \times 1$ vector of the response variable, X_0 is a unit vector of size $n \times 1$ and $X_0, X_1, \ldots, X_k$ are regressors, $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters and ε is an $n \times 1$ vector of error terms. The OLS estimator of β, the model parameter is given by $\hat{\beta} = (X^T X)^{-1} X^T Y$, where $X = [X_0, X_1, \ldots, X_k]$.

b. KNN method

This algorithm searches the pattern space for the K- training tuples closest to the unknown tuple. The closeness is defined in terms of distance between the tuples. It is better to normalize the values of each attribute before computing the closeness. For KNN, the unknown tuple is assigned the average value of its K- nearest neighbours as the predicted value.

c. SVR method

Support Vector Regression (Smola, 2002) is based on Statistical learning theory (Vapnik, 1995). Consider a linear regression model: $f(x) = w^T X + b$, where w is the weight of vector, b is the bias and X is the input feature vector. Then a solution that minimizes the error function is

$$f(x) = \sum_{i=1}^{n} (a_i^* - a_i) X^T X + b$$

where a_i^* and a_i are lagrange multipliers. The non-zero lagrange multipliers based on training vectors are called support vectors. The model for nonlinear case based on kernel function can be represented as:

$$f(x) = \sum_{i=1}^{n} (a_i^* - a_i) k(X^T X) + b$$

The Gaussian kernel is commonly used kernel function in $k(.)$ in SVR.

d. RT method

Even though the RT is an efficient method to produce outcomes, the main disadvantage of the RT is that it assigns the same average value of the response variable belonging to a particular group as the predicted value (constant) for all the observations in a group. Since the original values of the response variable are not the same, even though RMSE is minimum, the predicted value of the response value is of great interest to the decision-making process. In RT, two branches are grown from each node N corresponding to the condition $X_i <$ splitting point and $X_i \geq$ splitting point respectively. The splitting variables and splitting point define the rectangles D_j as

$$\begin{aligned} D_1 &= \{X \mid X_i < \text{splitting point}\} \qquad i = 1 \text{ to } k \\ D_2 &= \{X \mid X_i \geq \text{splitting point}\} \end{aligned}$$

where X_i is the i-th predictor variable, k denotes number of predictor variables. Then predicted value at p-th node is given by

$$\hat{m}_p = \frac{\sum_h Y_h I\{h \in D_i\}}{|D_j|}$$

where $p = 1$ to m and $|D_j|$ denotes number of observations in p-th node. Thus, an estimate of $m(x)$ in D_j is simply the average response of the Y observations with predictor vector in D_j. The goal is to find that combination of splitting variables and splitting point, which leads to minimum residual sum of squares (RSS). In each node, fitted value of the response variable is constant, $\hat{m}_p$.

Proposed hybrid regression model (RT-MLR)

The formulation of proposed hybrid model is as follows: initially dataset splits into several branches based on the RT algorithm. Branching is depends on the splitting variables (significant variables) and best split point, which produces the minimum error. Using RT, the best subset of variables is selected and redundant features are eliminated. The dataset in each leaf node is arranged based on the position of tuples that satisfy the corresponding splitting criterion. Further, a MLR is built for each leaf node with significant variables. The model parameters are estimated using the least-squares method. Since observations within each group show high intra class similarities, the application of MLR in each group separately ensures that the estimated regression function fits well with the data. This hybrid model is easy, flexible and simplifies the work of selecting the best set of variables separately.

The workflow of the proposed model is as follows:

- Apply RT algorithm to train dataset to construct a RT which holds the split point, leaf node and significant variables.
- In each leaf node, datasets are arranged according to the positions of the observations, which satisfies the corresponding splitting criterion.
- Fit MLR model separately, obtain the fitted values and repeat this for all the leaf nodes.

Therefore, predicted values at p-th node is given by

$$\hat{Y}_{hp} = \hat{\alpha} + \hat{\beta} X_h \qquad\qquad h = 1 \text{ to } n_p$$

This model is comprises of two steps: significant variables selection using RT and applying MLR to each leaf node separately to get improved prediction results. Observe that in each node, fitted values of the response variable are not constant. Since outputs of RT will be used in MLR, the proposed model performs better irrespective of problems such as missing values, noise and outliers. The proposed model can be used to identify the significant variables and causal parameters.

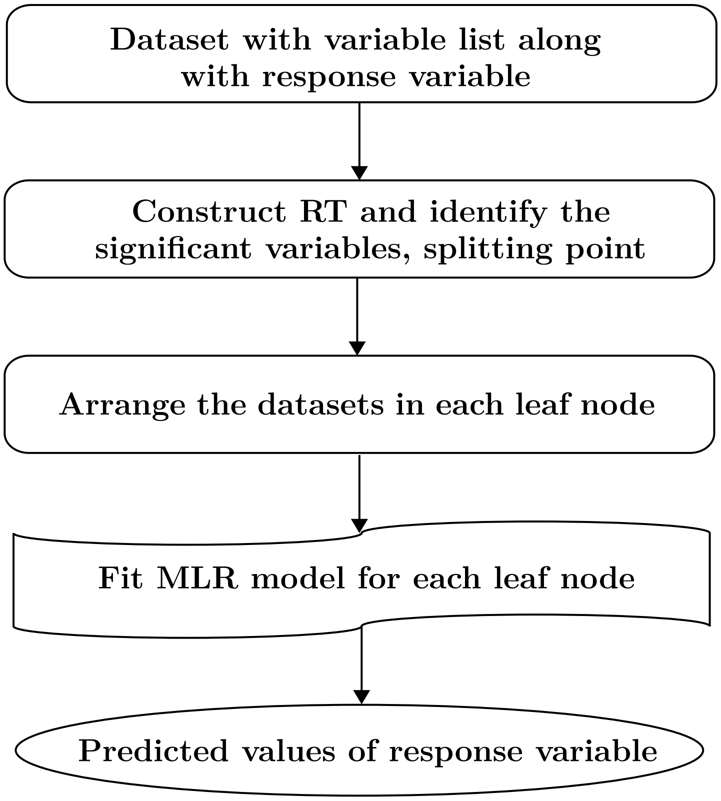
Dataset with variable list along with response variable
Construct RT and identify the significant variables, splitting point
Arrange the datasets in each leaf node
Fit MLR model for each leaf node
Predicted values of response variable

Figure 1: Flow chart of the proposed model

Performance measures

The model performance measure used in the simulation study and data analysis are

$$\begin{aligned}\text{Root mean square error (RMSE)} &= \sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}\\ \text{Mean Absolute Error (MSE)} &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}|Y_i-\hat{Y}_i|\\ \text{Coefficient of determination } (R^2) &= 1-\left[\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y}_i)^2}\right]\end{aligned}$$

3. Simulation study

In this section, a simulation study is performed to highlight the distinction between proposed hybrid RT-MLR model, RT, KNN regression, MLR model and SVR model. The predictive performance of these models is compared in terms of RMSE and MAE. The simulation design is as follows:

1. Considered the linear regression model $Y = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$, where all β's are set to 1 and X_1, X_2, X_3 are generated randomly from a standard normal distribution.

2. X_1, X_2, X_3 are also generated randomly from a uniform distribution $(0, 1)$ to check the robustness of the proposed model.

3. The error ϵ is generated from normal distribution with mean $= 0$ and variance $= 5$. The samples size used are $20, 50, 80, 100, 200, 500, 800, 1000, 2000, 3000$. The tree was grown to consist of three leaf nodes. The threshold stopping parameter for the RT is chosen as 0.01. For each scenario, 5000 repetitions were performed. In each simulation, the model is constructed using a train set and performance is evaluated using independently generated test data.

Table 1: Performance of different regression models for different sample size when the observations are from a standard normal distribution

Sample size	Method	RT	KNN	MLR	RT-MLR	SVR
20	RMSE	19.6	15.69	4.59	3.74	10.47
	MAE	15.65	12.19	3.75	3.15	8.3
50	RMSE	15.82	12.24	4.78	3.85	9.36
	MAE	12.4	9.38	3.83	3.15	7.45
80	RMSE	13.77	11.01	4.87	3.81	9.08
	MAE	10.72	8.44	3.9	3.12	7.26
100	RMSE	12.92	10.39	4.87	3.82	8.64
	MAE	10.05	7.98	3.9	3.12	6.9
200	RMSE	11.84	9.08	4.94	3.9	8.12
	MAE	9.27	7.02	3.94	3.18	6.47
500	RMSE	12.74	7.94	4.97	4.26	7.35
	MAE	10.04	6.18	3.97	3.45	5.62
800	RMSE	13.29	7.49	4.97	4.6	7.02
	MAE	10.47	5.86	3.97	3.71	5.45
1000	RMSE	13.59	7.33	5	4.71	6.78
	MAE	10.71	5.75	3.98	3.76	5.18
3000	RMSE	14.61	6.58	4.92	4.94	6.23
	MAE	11.52	5.24	3.91	4.04	4.84

From Table 1, the proposed hybrid RT-MLR model outperforms all other with a significant margin irrespective of sample size. The proposed model, along with overcoming the disadvantage of the regression tree, also performs better than all other models.

Robustness of the proposed hybrid RT-MLR model

To check the property of robustness of the proposed model about distributions, observations are generated from uniform distribution and results are summarised in Table 2.

Observe that proposed hybrid model outperforms all other models, irrespective of sample size.

Table 2: Performance of different regression model for different sample size when the observations are from uniform distribution

Sample size	Method	RT	KNN	MLR	RT-MLR	SVR
20	RMSE	7.92	9.22	6.14	4.97	7.76
	MAE	6.26	7.32	4.94	4.03	5.34
50	RMSE	7.26	8.83	6.67	5.04	8.74
	MAE	5.64	6.96	5.35	4.08	6.27
80	RMSE	7.02	8.74	6.8	5.06	9.12
	MAE	5.43	6.9	5.45	4.09	6.74
100	RMSE	6.88	8.65	6.83	5.13	9.26
	MAE	5.34	6.83	5.46	4.15	6.82
200	RMSE	6.85	8.51	6.93	5.08	9.56
	MAE	5.34	6.72	5.53	4.1	7.12
500	RMSE	7.27	8.39	6.98	5.44	9.85
	MAE	5.69	6.62	5.58	4.38	7.5
800	RMSE	7.51	8.31	6.98	5.71	9.89
	MAE	5.89	6.57	5.58	4.59	7.62
1000	RMSE	7.62	8.32	6.99	5.8	9.94
	MAE	5.97	6.56	5.58	4.66	7.69
3000	RMSE	8	8.24	7.01	6.28	10.1
	MAE	6.26	6.51	5.59	5.03	7.84

4. Real life application

The dataset used for the analysis consists of 96 observations and 12 variables related to the prices of cashew nuts collected from Dakshina Kannada. The variables considered are raw length, raw breadth, raw thickness, raw width, kernel length, kernel breadth, kernel thickness, kernel width, net count, sinkers count, moisture, out turn and price of the kernel. The price of the kernel is calculated based on the quality of the kernel obtained after deshelling the raw cashew nuts. In this study price of the kernel is taken as the response variable. Initially, a sample of 5 kg was drawn from each lot at different spots and then by hand halving method, a final sample of 1kg was drawn from these samples. Similarly, 96 such representative samples were drawn. For each sample of 1 kg, measurements on 12 variables mentioned is recorded. The dataset is randomly split into training and testing data sets in a ratio of 70:30. Each experiment is repeated five times with randomly selected test sets and train sets. The average performance over 5-fold validation is reported in Table 3. The performance of different regression models RT, KNN, MLR, SVR and hybrid RT-MLR model were recorded. In the analysis, we used most of the default arguments present in the packages.

Table 3: Performance measures for different regression models on test set

Regression model	RMSE	R^2
RT	29.24	58.57
KNN	26.70	68.67
MLR	33.23	45.24
SVR	30.52	62.65
RT-MLR	14.49	88.15

Table 3 shows that proposed hybrid RT-MLR model outperforms all other regression models with a significant margin based on RMSE and R^2. Thus, the proposed model can be used as an effective tool to fetch the most appropriate price of cashew nuts.

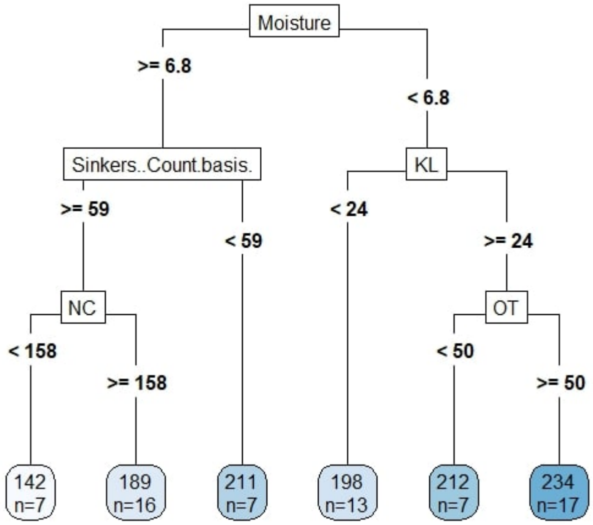

Figure 2: RT-MLR hybrid tree for cashew nuts price prediction

The above hybrid RT-MLR model suggests that the most important predictor variable for cashew nuts price is moisture content in the raw cashew nuts. Also proposed model inherently searches the interaction effects, as seen above. The interpretation of these effects is straightforward.

According to Figure 2, if Moisture $<$ **6.8** and Kernel length $\geq$ **24** $\Longrightarrow$ *High price* and if Moisture $\geq$ **6.8** and Sinkers count $<$ **59** $\Longrightarrow$ *High price.*

The proposed hybrid RT-MLR model is used to predict the cashew nut prices and to identify important casual variables and relationships. To get the optimal price for the cashew nuts, proposed model recommends checking the moisture level, kernel length and sinkers count and decide the price for the cashew nuts as shown above. Since out turn cannot be controlled at the time of purchase or selling, only controllable parameters are given based on the regression analysis performed using the hybrid model. The hybrid model along with improved accuracy, also helped the buyers and sellers to decide the most reasonable price for the cashew nut.

5. Conclusion

The main objective of this paper is to develop a hybrid model for improving prediction accuracy. This paper proposes a hybrid regression model based on the RT and MLR model. The proposed model also successfully overcomes one of the main disadvantage of the RT. The prediction performance of the proposed hybrid model is compared with many popular regression models through a simulation study. The simulation results indicate that the proposed hybrid model outperforms all other models when observations are generated from a normal distribution. The simulation results also demonstrate that the proposed hybrid model is fairly robust. The empirical study shows that hybrid model helped the cashew nuts buyers and farmers to decide the most appropriate price for the cashew nuts with improved prediction accuracy. The main advantage of this model is its easy interpretability. The proposed hybrid RT-MLR model can be used for handling both linear and non-linear datasets effectively. The proposed model approach is extended for classification problems as future work.

References

Bennett, K. P. and Blue, J. (1998). A support vector machine approach to decision trees, *In: 1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227)*, **3**, 2396–2401.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees.* CRC Press, Boca Raton.

Chakraborty, T., Chakraborty, A. K., and Mansoor, Z. (2019). A hybrid regression model for water quality prediction. *OPSEARCH*, **56**, 1167–1178.

Chakraborty, T., Chakraborty, A. K., and Chattopadhyay, S. (2019). A novel distribution-free hybrid regression model for manufacturing process efficiency improvement. *Journal of computational and Applied Mathematics*, **362**, 130–142.

Chakraborty, T., Chattopadhyay, S., and Chakraborty, A. K. (2018). A novel hybridization of classification trees and artificial neural networks for selection of students in a business school. *Opsearch*, **55**, 434–446

Chang, F. and Liu, C. C. (2012). Decision tree as an accelerator for support vector machines. *In: Ding, X. (ed.) Advances in Character Recognition. IntechOpen*, Rijeka.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**, 273–297

Kumar, M. A. and Gopal, M. (2010). A hybrid SVM based decision tree. *Pattern Recognition*, **43**, 3977–3987.

Lola, M. S., Ramlee, M. N. A., Gunalan, G. S., Zainuddin, N. H., Zakariya, R., Idris, M. S., and Khalil, I. (2016). Improved the Prediction of Multiple Linear Regression Model Performance Using the Hybrid Approach: A Case Study of Chlorophyll-a at the Offshore Kuala Terengganu. *Open Journal of Statistics*, **6**, 789–804.

Mienyea, I. D., Suna, Y., and Wangb, Z. (2019). Prediction performance of improved decision tree-based algorithms: a review, *2nd International Conference on Sustainable Materials Processing and Manufacturing.*

Scholkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, Massachusetts.