ORIGINAL RESEARCH PAPER
Volume-6 | Issue-11 | November-2017 | ISSN No 2277 - 8179 | IF : 4.176 | IC Value : 78.46
INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

IMPROVEMENT IN RANDOM FOREST REGRESSION MODEL USING A GROWTH PREDICTOR

Computer Science

Bhavna Saluja WizeCommerce
Simmi Saluja Maharaja Agrasen Institute of Technology

ABSTRACT

Customer data analytics refers to the systematic study of the company's customer information that helps make key business decisions. By market segmentation and predictive analysis, we can identify and retain the profitable customers, locate sites for a successful business and understand customer relationship management. Analysing user intentionality towards products is crucial for retargeting. Predictive models give a numerical measure which is linked with each client for their propensity to churn and the result is in terms of probability. This information proves to be very useful in developing marketing campaigns aimed at customer retention. The research in this work is based on the empirical results obtained by making regression models on consumer retail data sets. Random Forest algorithm is one of the best algorithms that is used today for this purpose. In this paper, an attempt has been made to improve the accuracy of the model through introduction of a metric- growth variable- associated with predictor variables. It is also discussed how coverage or accuracy of the Random Forest model can be improved by careful selection of the variables being used to build the model and looking at the data from different perspectives and summarizing the relationships identified. Experimental results clearly indicate the enhancement in Random Forest model in terms of accuracy.

KEYWORDS

Data Mining, Random Forest, Tree Performance, Decision Trees, Regression, Predictor

Customer churn is defined as the probability measure value of customers to leave a firm in given time period. Higher churn rate implies higher chances of a customer's proclivity leaving the firm. Churn rate, also called attrition rate, is a popular topic today to investigate for many businesses and industries. It has become crucial to inspect the reasons for customers losing their interests in products or choosing other competitors over them. Machine learning and data mining are the two versatile and powerful methods to assess the effect of different predictors on the churn rate. Churn rate is especially significant for subscription-based firms such as telecommunication companies. A customer stays or leave is the churn rate prediction which comes under the field of classification problem.

CART analysis is an exploratory research method for describing relationships between variables (Kuhn, Page, Ward, & Worall-Carter, 2013). Improving accuracy in classification tasks is the topic of limelight today in data mining. In past few years, the topic has been grasping attention from many researchers all over the globe. The greater the number of suitable classification methods, more difficult it is to find the most effective one for our application and which parameters to use for its validation. A number of studies using various algorithms, such as sequential patterns, classification trees, SVM and neural networks, have been conducted for exploratory analysis of customer churn and to show the prospects of data mining through experiments and case studies(Xie, Li, Ngai, & Ying, 2009).

A new algorithm was proposed by Leo Breiman in 2001, known as Random Forest, which has come upon as a new scheme towards exploration and analysis of data (Breiman, 2001). Multiple decision trees are formed by creating models with forecasting probabilities. The randomization in Random Forest algorithm occurs in such a way that random samples of data are taken for bootstrap samples, and then input attributes are randomly selected to form individual base decision trees (Kulkarni & Sinha, 2014).

In this paper, we have attributed our research to improve the accuracy of the Random Forest model via proposing a new metric– growth factor – which can be associated with selective variables. This paper puts forth certain observations regarding the selection of number of decision trees, picking up the optimized sample size and discarding the negatively impacting variables.

Random forest (RF) is an ensemble classification approach that has proved its high accuracy and superiority. Random Forest has gained considerable attention from the research community. (Fawagreh, Gaber, & elyan)
Algorithm(Edwards & Gaber, 2014):
Input N, S
{Process}
Create an empty vector $\overrightarrow{RF}$

for i=1 → N do

Create an empty tree T_i repeat
Sample S out of all features F using Bootstrap sampling
Create a vector of all features F using Bootstrap sampling

Create a vector of the S features $\overrightarrow{Fs}$

Find Best Split feature B ($\overrightarrow{Fs}$)

Create a New Node using B ($\overrightarrow{Fs}$) until No More Instances to Split On

Add Ti to the $\overrightarrow{RF}$ end for {Output}

A vector of trees $\overrightarrow{RF}$.

The random forests algorithm (for both classification and regression) is as follows:

- The original data is split into n-tree bootstrap samples.
- Best split needs to be chosen. For selection of the same, an unpruned classification or regression tree is grown considering the bootstrap samples. M-try predictors are sampled randomly and the best among them is chosen.
- Aggregation of the prediction results of n-trees are considered to predict new test data(Liaw & Wiener, 2002).

Random forests uniqueness lies in the way it creates classification and regression trees after constructing each tree using a different bootstrap sample of the data. The standard way of splitting each node of a tree is to choose the best split among all variables. But, the approach used by random forest to split a node selects the best among a subset of predictors randomly chosen at that node. This new way of splitting a node tends to show better performance in comparison to other classifiers, including discriminant analysis, support vector machines and neural networks, and is robust against overfitting (Liaw & Wiener, 2002). Moreover, this new strategy is user-friendly as it inculcates only two parameters (number of trees in the forest and number of variables in the random subset at each node).

Additional randomness was introduced with it because of the formation of decision trees by classification and regression trees (CART) technique. Gini index heuristics are used to examine the subset of features at each node using this technique. Higher Gini index value helps to choose the feature which acts as split feature for that node. Gini index has been introduced by Breiman et al. (1984). However, it has been first introduced by the Italian statistician Corrado Gini in 1912. The index is a function that is used to measure the impurity of data, that is, how uncertain we are if an event will occur (Corchado, Yin, Quintián, & Lozano, 2014). It is defined as given below:

$$\text{Gini}(t) = 1 - \sum_{i-1}^{N} P(C_i|t)^2$$

where t is a condition, N the number of classes in the data set, and C_i is the i^{th} class label in the data set (Corchado, Yin, Quintián, & Lozano, 2014). Our approach in this review paper is to show how the accuracy of this successful classification technique can be improved.

Additional relevant information is provided by the Random Forest package. One of them is to estimate the importance of the predictor variables, and the other assesses the internal structure of the data (the proximity of different data points to one another) (Liaw & Wiener, 2002).

1) **Variable Importance:**
The importance of a variable is determined by the random forest algorithm by looking at the rate of increase of prediction error when permutations of (OOB) data for that variable are taken, keeping all other variables unchanged. While the construction of random forest is in progress, calculations are done tree by tree. Variable Importance parameter is a difficult to explain in general, as the importance of a variable can arise due to its (possibly complex) interaction with other variables.
2) **Proximity Measure:**
The (i, j) element of the proximity matrix produced by Random Forest is defined as the fraction of trees in which elements i and j fall in the same terminal node. Similar observations should lie in the same terminal nodes more often than dissimilar ones. The proximity matrix is useful for identification of structure in the data or for unsupervised learning(Liaw & Wiener, 2002).

METHOD:
Regression analysis is a statistical procedure to estimate the associations among variables. It encompasses various techniques and strategies for modelling and analysing several variables, to derive the relationship between a dependent variable and one or more independent variables (or 'predictors'). In brevity, regression analysis assists in understanding as to how the dependent variable (or 'criterion variable') changes due to the variation in any one of the independent variables, when other independent variables are kept fixed (Wikipedia, 2016).

A new variable is introduced in the regression equation of random forest model which enhances the performance of the model generated in terms of accuracy & precision.

Classical Regression equation:

$$y \sim x_1 + x_2 + \cdots + x_n \quad (1)$$

where, y = predicted response, x_i = predictor variables
Proposed Regression equation:

$$y \sim x_i + Gr_i \quad (2)$$

where, y = predicted response, x_i = predictor variables, Gr_i = Growth variables
(adding growth predictor associated with other predictor variables)

$$Growth(i) = \frac{F(i)-I(i)}{I(i)} = \frac{F(i)}{I(i)} - 1 \quad (3)$$

where F(I) = Final value of i^{th} predictor, I (i) = Initial value of i^{th} predictor

Growth variable: It is the growth factor associated with the corresponding predictor scaled according to the range of values obtained.

For instance, suppose 'revenue' is one of the predictors of a customer's purchasing behaviour, then if revenue for a customer is defined from time, t=1 to time, t=n of the dataset, we calculate the growth factor of revenue associated with that customer as follows –

$$Gr(revenue) in\ time\ n = \left(\frac{Revenue\ in\ time\ n}{Revenue\ in\ time\ 1}\right)^{\frac{1}{n-1}} - 1 \quad (4)$$

Scaling the growth exponentially by (1/ (n-1)) is done because the numbers corresponding to the usual growth factor ((F-I)/I) range from a very small value to a very large value in most of the datasets.

Following data sets were considered for the experiment:
Data Set 1: Online retail dataset is a transnational data set containing transactions of around 7 months for a UK-based non-store online retail (Chen, Sain, & Guo, 2012)(Archive.ics.uci.edu, n.d.).

Data Set 2: FoodMart dataset is a sample data set taken from Microsoft containing market baskets for around 8000 consumers and over 1500 products. The data set of aggregate sales for 1997 was used for the analysis(Recsyswiki.com, 2016).

The predictive modelling training covers detailed steps - target variable definition, data preparation & treatment and model development. The implementation of the model is done in R.
Following tasks were done before performing the experiments:

Data Preprocessing (cs.ccsu.edu): Data cleaning was done by populating missing values in accordance with the other values, filtering the noise in the data, and rectifying inconsistent data using domain knowledge. Afterwards, data reduction was performed which comprised of decreasing the volume while maintaining similar results. In the end, data discretization that forms a part of data reduction, replaced the numerical attributes with nominal ones.

Variables Selection: In this work, the datasets were categorized on per customer per month/quarter basis. Separate files were formed for different times to find the churn rate of customers considering their purchasing pattern for the time period observed.

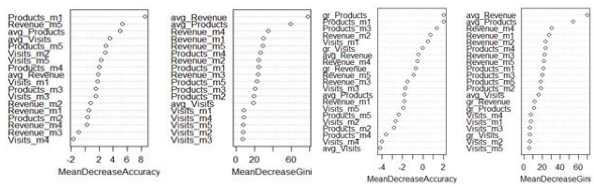
For data set-1, the month of May 2011 has been kept as processing time and the month of June 2011 has been used as the target data (whether a customer is a churner or a non-churner considering their data of previous months). For data set-2, the quarter 3 of the year 1997 has been kept as processing time of the model.

Three new variables were added to the dataset - namely, revenue generated by a customer in a month, no. of products purchased by a customer in a month/quarter, and no. of visits a customer has made in a month/quarter – which act as the attributes for the predictive modelling.

A new growth variable was linked with revenue, no. of products purchased, no. of visits of a customer and an improvement was observed in the coverage of the predictive model developed using random forest.

The inverted target variable was passed in the model which considers that the churners are denoted by 1 (positive cases for the model) and non-churners are denoted by 0 (negative cases for the model).

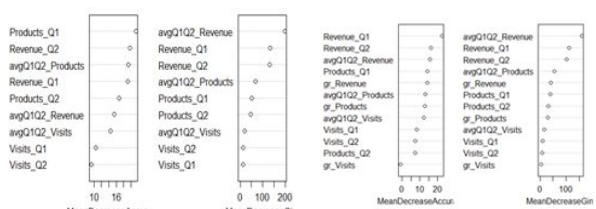
RESULTS:
I. Dataset 1: Online Retail



Without growth predictor With growth predictor

Fig. 1. Variable importance plots of online retail dataset

II. Dataset 2: FoodMart



Without growth predictor With growth predictor

Fig. 2. Variable importance plots of foodmart dataset

Table 1. Performance of Random Forest Regression Model

Model's Output Parameters	Dataset1		Dataset2	
	Without Growth predictor (%)	With Growth predictor (%)	Without Growth predictor (%)	With Growth predictor (%)
Accuracy	53.26	55.88	58.75	60.00
Error Rate	46.74	44.12	41.25	40.00
Recall	32.48	35.24	37.83	60.77
Specificity	71.94	74.42	75.15	59.01
Precision	51.00	55.33	54.42	55.05

DISCUSSION:

The experimental results, for different datasets, as given in Table 1 indicate that: the proposed new predictor variable in the regression equation not only improves the accuracy of the model but increases the recall and precision as well. These results verify the effectiveness and rationality of the proposed feature when it is used in regression models. This paper presented a novel variable selection in regression equation for predictive modelling using Random Forest algorithm. To show the usefulness and effectiveness of the new variable - Growth predictor, experiments were conducted on two different consumer retail datasets for building predictive models using the Random Forest algorithm. The experimental results show that the model built without the growth predictor has a lower coverage than that of the model built with the growth predictor. Since the coverage rate of the model with growth predictor is higher, it proves to be a more accurate prediction model, and hence the response given by it is more reliable. Thus, the addition of growth predictor improved the performance of random forest model by increasing the accuracy and precision of the model.

REFERENCES:

1. Kuhn, L., Page, K., Ward, J., & Worall-Carter, L. (2013, November). The process and utility of classification and regression tree methodology in nursing research. *Journal of Advanced Nursing*.
2. Xie, Y., Li, X., Ngai, E., & Ying, W. (2009, April). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications, 36*(3), 5445-5449.
3. Breiman, L. (2001, October). Random Forests. *Machine Learning, 45*(1), 5-32.
4. Kulkarni, V. Y., & Sinha, P. K. (2014, May). Effective Learning and Classification using Random Forest Algorithm. *International Journal of Engineering and Innovative Technology, 3*(11).
5. Fawagreh, K., Gaber, M. M., & elyan, E. (n.d.). Random forests; from early developments to recent advancements. *Systems Science and Control Engineering, 2*(1).
6. Edwards, K., & Gaber, M. M. (2014). *Astronomy and Big Data.* Springer International Publishing.
7. Liaw, A., & Wiener, M. (2002, December). Classification and Regression by randomForest. *R News, 2/3*, pp. 18-22.
8. Corchado, E., Yin, H., Quintián, H., & Lozano, J. A. (2014). Intelligent Data Engineering and Automated Learning. *IDEAL: International Conference on Intelligent Data Engineering and Automated Learning.* Salamanca: Springer.
9. Wikipedia. (2016). *Regression Analysis.* Retrieved from https://en.wikipedia.org/: https://en.wikipedia.org/wiki/Regression_analysis
10. Archive.ics.uci.edu. (n.d.). *UCI Machine Learning Repository: Online Retail Data Set*. Retrieved from https://Archive.ics.uci.edu: https://Archive.ics.uci.edu/ml/datasets/Online+Retail#
11. Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing and Customer Strategy Management, 19*, 197-208.
12. Recsyswiki.com. (2016). *Category: Dataset- RecSysWiki.* Retrieved from https://recsywiki.com/: https://recsywiki.com/wiki/Category:Dataset
13. cs.ccsu.edu. (n.d.). *Data Preprocessing.* Retrieved from http://www.cs.ccsu.edu/: http://www.cs.ccsu.edu/~markov/ccsu_courses/datamining-3.html