Supplier Segmentation using boosting tree – A Machine Learning approach

Mrugank Akarte
Production Engineering Department
Vishwakarma Institute of Technology, Pune, India
mrugankakarte13@gmail.com

Prof Priyanka Verma,
Assistant Professor (Industrial Engineering and Manufacturing Systems)
National Institute of Industrial Engineering (NITIE)
Vihar Lake, Mumbai, India, 400 087
priyankaverma@nitie.ac.in

***Abstract*- Supplier segmentation play a crucial role in the decision-making process to maximize business performance. Previous research shows that a supplier score was obtained based on multiple parameters and then this score was used to classify them. This paper presents a machine learning approach for supplier segmentation based on the supplier capabilities and willingness criteria ratings. Dataset developed by Jafar Rezaei and Roland Ortt is used as training set, and cross-validation error is used as a performance metric for the model. The hyperparameters are tuned using 3-Fold cross-validation. The future supplier can be segmented using this model based on their capabilities and willingness criteria ratings.**

***Keywords*: Supplier segmentation, Classification, machine learning**

1. INTRODUCTION

Today purchasing accounts for over 50 percent of the cost of goods sold[1]. Realizing this increased in spend firms were forced to change the role of purchasing function from traditional clerical or administrative to the strategic[2]. Strategic procurement plays a vital role in achieving profitability and hence the competitiveness of an organization. Increased globalization, technological advances, a shift towards outsourcing, and changing customer patterns are some of the key factors influencing the purchasing strategies.

Supplier segmentation is the strategic decision-making process of a firm where the supply base is classified after the selection of a supplier into different groups. It helps organizations in applying different supplier management strategies for each group to maximize business performance. A systematic approach to classify suppliers helps in developing a sound strategy by minimizing the risk and leveraging the supplier's capabilities for competitive advantage.

Supplier selection and then classification has been widely addressed by the researchers including the multi-criteria decision making where each supplier score is obtained on multiple performance parameters and then based on the score obtained they are classified into various groups such as excellent, very good, good, etc. This work focuses on the use of a machine learning approach, especially boosting tree in segmenting the suppliers. Relevant literature has been reviewed and is briefly presented next followed by the methodology adopted in supplier segmentation, results, and conclusion.

2. LITERATURE REVIEW

In today's competitive environment, organizations are excelling based on their supply chain. Supplier plays a vital role in the competitiveness of an organization whether it is a lean or agile supply chain. Supplier selection has been described as one of the critical processes in the purchasing and supply management function, and the problem has received significant attention in the literature including the review from time to time. For more information on the criteria and tools used in supplier selection, one can refer to the recent reviews on supplier selection[3]. However, supplier segmentation has received less attention even though it is an essential decision in deciding the strategy in supplier management. A brief description of supplier segmentation including the definitions and related literature is given next.

Supplier segmentation means categorizing suppliers of a specific firm based on their similarities to arrive at a manageable number of segments and each of which requires a separate strategy[4]. Supplier capabilities and willingness have been used for segmenting the suppliers. Jafar Rezaei, et al. (2013)[4][5] defined supplier segmentation after analyzing various criteria's like price, quality, delivery, reserve capacity, geographical location, financial position, commitment to quality, communication openness, willingness to share information and long-term relationship for supplier selection process. Further, the authors defined supplier capabilities, and supplier willingness as follows:

"Supplier segmentation is the identification of the capabilities and willingness of suppliers by a particular buyer in order for the buyer to engage in a strategic and effective partnership with the suppliers with regard to a set of evolving business functions and activities in the supply chain management."

"*Supplier's capabilities are complex bundles of skills and accumulated knowledge, exercised through organizational processes that enable firms to coordinate activities and make use of their assets in different business functions that are important for a buyer*."

"*Supplier's willingness is confidence, commitment and motivation to engage in a (long-term) relationship with a buyer.*"

The suppliers can also be segmented based on their historical performance. However, the approach does not help in the strategic decision-making process as it lacks in important criteria and the information required (Example, willingness to share information and ability to communicate openly) for the same in the decision-making process. Given this, researchers have identified the necessary criteria through an in-depth literature review and proposed systematic approaches to segmenting the suppliers. These are briefly given next.

Organizations use different approaches to manage the supplier relationship. Dyer, Jeffrey H, et al. (1998)[6] studied supplier management approaches across different manufacturers and countries and highlighted the strength and weaknesses of each approach.

Rezaei and Ortt proposed two approaches for segmentation of the suppliers in four groups based on two dimensions namely Capabilities and willingness of the suppliers. Capabilities are measured by using six criteria (Price, Delivery, Quality, Reserve capacity, Geographical location, and Financial position), and willingness is measured by six criteria (Commitment to quality, Communication openness, Reciprocal arrangement, Willingness to share information, Supplier's effort in promoting JIT principles, Long-term relationship). In the first approach, the authors used fuzzy rule-based systems where the degree of supplier capabilities and willingness was aggregated in a single output for the above-listed criteria. In another approach, the authors proposed a fuzzy Analytic Hierarchy Process (AHP) for segmentation of the suppliers. The proposed methodology is used to address the ambiguities and uncertainties that usually exist in human judgment and is demonstrated with an example for a broiler company to classify 43 suppliers into four segments. Based on the classification, the authors further suggested strategies to handle different segments. For example, suppliers that have low capabilities and low willingness are referred to as type 1 segment suppliers that buyer can avoid or replace.

Manoj Hudnurkar, et al. (2016)[7] conducted exploratory research and identified 26 criteria used for supplier classification based on the interviews of practitioners from buyer multinational manufacturing companies operating in India. However, authors reported that the majority of these criteria resemble the factors reported in prior research.

Marina Segura and Concepción Maroto (2017)[8] proposed a multiple criteria system for systematic decision making for supplier segmentation that considers group decision making and internal as well as external information (both, quantitative and qualitative) about products and provider. The system has been validated in a real implementation in a big manufacturing company, which works for several sectors such as food, pharmaceuticals, and chemicals.

The brief review of the literature indicates that supplier segmentation is very critical in supply chain management and competitiveness of an organization and many criteria have been identified in selecting and segmenting the suppliers. Also, multi-criteria decision-making approach is the most commonly adopted methodology in supplier selection and classification. However, to the best of our knowledge, there is no reported work on the use of boosting tree – a machine learning approach – in the segmentation of supplier.

This work presents the application of a machine learning approach (boosting tree) for supplier segmentation decision making the process by taking an example from the literature of Jafar Rezaei and Roland Ortt. The methodology used, and results obtained are briefly given next.

REVIEW OF METHODS

The following section introduces the XGBoost algorithm, Cross-validation and Multiclass logloss error.

A. XGBoost

XGBoost is a scalable end-to-end machine learning system for tree boosting, which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges[9]. Maksims Volkovs, Guangwei Yu and Tomi Poutanen won the 2017 ACM RecSys Challenge where the task was, given a job posting, the goal was to identify the user that may be interested in receiving the job posting as a recommendation and also to identify the users that are appropriate for the given job. The models were trained using XGBoost library and outperformed other solutions consistantly[10]. XGBoost has been used for physics experiment to identify a rare decay phenomenon in Large

Hadron Collider, search result relevance, context click ads and many more. The details of use cases are available on the official repository page of XGBoost library[11]. It is an open source library available in C++, R, Python, and Julia.

XGBoost is faster and more efficient implementation of boosting algorithm. The idea behind the boosting algorithm is to combine weak learners into a strong learner in an iterative manner. Initially, a model is fit to data; then another model is fit to the residuals, which corrects the error from the previous model. A new model is then created by combining the original model and a model fit to the residuals. This process is repeated until a strong learner is obtained. The classifier designed for this problem outputs the probability that supplier can be classified in a given segment.

B. k-Fold Stratified Cross-validation

k-Fold Cross-validation is a technique for measuring how well the machine learning model will perform on an independent test set[12]. The training data set is divided into k sets, and out of k sets (k-1) sets are used for training the algorithm, and the left-out set is used as a validation set. This process is repeated until every set has the chance to be a validation set. Stratified sampling ensures the same proportion of observations across each category in every set.

In this problem, 3-Fold stratified cross-validation is used. The error is calculated by averaging the multiclass logloss error across all the three folds. In other words, first fold 1 and fold 2 are used as training set and fold 3 is used as a validation set, then the error on fold 3 is calculated. This process is repeated with fold 1 and fold 2 as a validation set. Finally, the average error across all the three folds is calculated which is the measure of performance for the model. This is also referred to as cross-validation error and the error measured on the training set is known as training error.

C. Multiclass Logloss

Logloss is a loss function which measures the performance of a classification model where predictions have probability values between 0 and 1. Log loss increases as the prediction diverge from the actual result. For example, predicting a probability of 0.05 for the actual value of 1 will result in high logloss. Logloss for binary classes is formulated as shown below.

$$logloss = -\frac{1}{N}\sum[y \log Pc + (1-y)\log(1-Pc)]$$

Where,

N = number of training examples,
y = actual value (0 or 1),
Pc = probability of observation equal to 1.

Generalizing above equation for more than two outcome classes, we get the following formula.

$$mlogloss = -\sum_{c=1}^{n} \mathrm{y} * \log(Pc)$$

Where,

c = outcome class number,
y = 1 when observation *o* belongs to class *c*, else it is equal to 0,
Pc = probability of observation *o* belonging to class *c*.

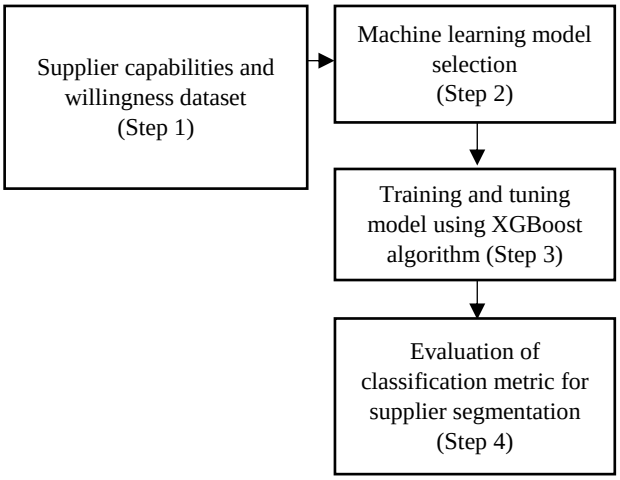
In our context, c will have values 1,2,3 and 4 for Type 1, Type 2, Type 3 and Type 4 segment respectively which are defined in the following section. Y is the variable which will hold the true labels of the observations, for example, if the true label of observation is 2, then y will be (0,1,0,0). Pc will have probability of the observations belonging to each segment. The error is then calculated across all classes after elementwise multiplication between y and log(Pc).

4. METHODOLOGY

The overall methodology for supplier segmentation using XGBoost algorithm is shown in figure 1. The important steps in this methodology are as follows:

- Step 1: Dataset developed by Jafar Rezaei and Roland Ortt using supplier capabilities and willingness is used as training data.
- Step II: XGBoost algorithm is used to determine the probabilities of supplier belonging to each class.
- Step III: Cross-validation is used to tune and train the model.
- Step IV: Classification metric was evaluated for the above approach using Multiclass Logloss error.

Figure 1: Overall Methodology for supplier segmentation



The dataset consists of 43 suppliers with 12 columns, 6 columns each representing different capabilities and willingness criteria as developed by Jafar Rezaei and Roland Ortt for supplier segmentation. The six capabilities and six willingness criteria are price (C1), delivery (C2), quality (C3), reserve capacity (C4), geographical location (C5), financial position (C6) and Commitment to quality (W1), communication openness (W2), reciprocal arrangement (W3), willingness to share information (W4), supplier's efforts in promoting JIT principles (W5), long-term relationship (W6) respectively. Since the size of dataset is very small, simple gradient boosting tree algorithm XGBoost is used to determine the effectiveness of machine learning in classifying suppliers into four different segments namely Type-1, Type- 2, Type- 3 and Type-4, where Type-1 refers to low capability-low willingness, Type-2 refers to low capability-high willingness, Type-3 refers to high capability-low willingness and Type-4 refers to high capability-high willingness.

5. DESIGN OF CLASSIFIER

XGBoost is used to train the model by minimizing the multiclass logloss error (mlogloss). The hyperparameters *eta*, *depth*, *lambda*, *subsample*, *colsample_bytree*, and *nrounds* were tuned using 3-fold stratified cross-validation. Here, *eta* is the 'learning rate' used by the algorithm, *depth* is the maximum depth of each tree, *lambda* is L2 regularization parameter on weights, *subsample* is the ratio of training instances to be used during training, *colsample_bytree* is ration of columns when constructing each tree and *nrounds* is number of trees to grows. The values for hyperparameters used in cross-validation as tabulated in table 1.

Table 1: Parameter values for cross-validation

Hyperparameters	Range of values used
eta	0.01, 0.03, 0.05, 0.07, 0.09, 0.1, 0.2, 0.25, 0.3, 0.35, 0.4
depth	3,4,5,6
lambda	0,1
subsample	0.5, 0.8, 0.9
colsample_bytree	0.5, 0.8, 0.9

The model was tuned for every combination of values mentioned in Table1. The optimal values for the model were found to be *eta = 0.3*, *depth = 5*, *lambda = 0*, *subsample = 0.9*, *colsample_bytree = 0.5* and *nrounds = 46*. Another parameter '*early_stopping_round*,' was set to 10, to prevent overfitting. This parameter prevents further training of model if the error function does not decrease within the given limit, in this case after 10 iterations.

The distribution of data-points in each fold after stratified sampling is mentioned in table 2. Figure 2 shows the distribution of data points from each segment into the folds.

Table 2: Distribution of data in Cross-validation

Fold	No of Suppliers	Supplier Id
Fold 1	14	2,4,9,11,12,18,22,27,31,32,34,35, 42,43
Fold 2	14	5,8,10,15,17,20,23,24,25,29,33,36 ,37,40
Fold 3	15	1,3,6,7,13,14,16,19,21,26,28,30,3 8,39,41

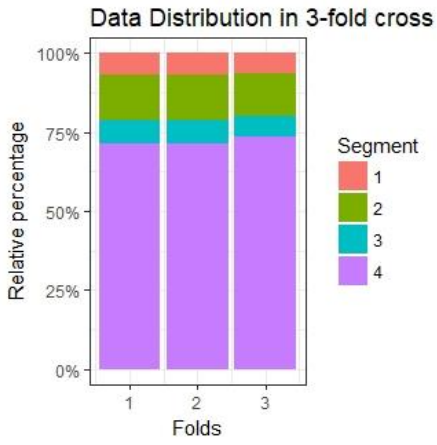


Figure 2: Relative percentage distribution of each segment in folds.

6. RESULTS

Since the size of the dataset is very small, it is not possible to divide the dataset into two sets, training, and validation. Hence, cross-validation error is used to determine the performance of the model. The cross-validation error for

the tuned parameters and a different number of training iterations is tabulated in table 3.

From the following table, it can be seen that validation error reduces significantly to iteration 46, this suggests that the training of the algorithm can be stopped earlier to prevent overfitting and avoid additional computational resources. For segmenting future suppliers, a model trained using optimal values mentioned above can be used.

Table 3: Classification error for the different number of iteration

Number of iterations	Training error	Cross-Validation error
6	0.323636	0.453150
26	0.093920	0.302212
46	0.087958	0.284169
66	0.086978	0.288486
86	0.086241	0.286920
100	0.085986	0.291749

Table 4 shows the predicted probabilities for suppliers in Fold 1 which is used as a validation set when the model is trained using Fold 2 and Fold 3. Similarly, predictions for Fold 2 and Fold 3 are obtained.

Table 4: Predictions of suppliers in Fold 1 used as a validation set (Green: Correct Classification, Red: Misclassified)

Supplier Id	Type 1	Type 2	Type 3	Type 4	Actual Segment
2	0.0053	0.0214	0.0076	0.9655	4
4	0.0008	0.0013	0.0014	0.9963	4
9	0.6506	0.2427	0.0998	0.0067	1
11	0.0008	0.0013	0.0014	0.9963	4
12	0.0074	0.0064	0.0086	0.9774	4
18	0.0644	0.7652	0.0179	0.1524	2
22	0.0044	0.0027	0.0021	0.9906	4
27	0.0009	0.0011	0.0016	0.9963	4
31	0.000993	0.001126	0.001727	0.996155	4
32	0.0009	0.0018	0.0016	0.9955	4
34	0.0944	0.1001	0.2199	0.5854	3
35	0.0010	0.0014	0.0018	0.9956	4
42	0.0015	0.0018	0.0020	0.9944	4
43	0.1978	0.5662	0.0377	0.1981	2

7. Conclusion

The paper presents a machine learning approach using boosting tree for supplier segmentation. In this work, an example of supplier segmentation from the literature has been used as an input to demonstrate the application of machine learning approach in supplier segmentation. The example consists of data of 43 suppliers for their 12 characteristics (classified into two groups namely capabilities and willingness) and their classification. The model, XGBoost algorithm, is then tuned using cross-validation. Cross-validation error is used as a performance matrix for the model. Also, the input dataset is unbalanced, that is one segment has more examples than the other. The model may get biased if the data points are small and unbalanced. This problem can be accounted in classification error to some extent by cross-validation approach. For segmenting future suppliers, a model trained using tuned parameters for 46 iterations can be used.

The machine learning tools can provide a better result if the larger dataset is available as an input. Machine learning approach simplifies the supplier segmentation process for an organization and helps to develop different supplier strategies for different segments by leveraging the supplier capabilities for competitive advantage.

8. Reference:

[1] Anna Duboisa, Ann-Charlott Pedersen, Why relationships do not fit into purchasing portfolio models - a comparison between the portfolio and industrial network approaches, European Journal of Purchasing & Supply Management 8 (2002) 35–42

[2] Mclvor RT, Humphreys PK and McAleen WE, A strategic model for the formulation of an effective make or buy decision, Management Decisions, 35/2, 1997, 169-178.

[3] Anton Wetzstein, Evi Hartmann, W.C. Benton jr., Nils-Ole Hohenstein, A systematic assessment of supplier selection literature – State-of-the-art and future scope, Int. J. Production Economics 182 (2016) 304–323

[4] Jafar Rezaei, Roland Ortt, Supplier segmentation using fuzzy logic, Industrial Marketing Management 42 (2013a) 507–517

[5] Jafar Rezaei and Roland Ortt, Multi-criteria supplier segmentation using a fuzzy preference relation based AHP, European Journal of Operational Research 225 (2013b) 75–84.

[6] Dyer, Jeffrey H, Dong Sung Cho, and Chu, Wujin, Strategic supplier segmentation: The next "best practice" in supply chain management California Management Review; winter 1998; 40, 2; pg. 57 -77.

[7] Manoj Hudnurkar, Urvashi Rathod and Suresh Kumar Jakhar, Multi-criteria decision framework for supplier classification in collaborative supply chain Buyer's

perspective, International Journal of Productivity and Performance Management Vol. 65 No. 5, 2016 pp. 622-640

[8] Marina Segura, Concepción Maroto, A multiple criteria supplier segmentation using outranking and value function methods, Expert Systems with Applications 69 (2017) 87–100

[9] Tianqi Chen, Carlos Guestrin, XGBoost: A Scalable Tree Boosting System, June 2016, available at: https://arxiv.org/abs/1603.02754 and accessed on July 27, 2018.

[10] Maksims Volkovs, Guang Wei Yu, and Tomi Poutanen. 2017. Content-based Neighbor Models for Cold Start in Recommender Systems. In Proceedings of RecSys Challenge '17, Como, Italy, August 27, 2017, 6 pages. https://doi.org/10.1145/3124791.3124792

[11] Distributed Machine Learning Community: XGboost https://github.com/dmlc/xgboost/tree/master/demo#usecases

[12] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, 'Resampling Methods' Introduction to Statistical Learning, Springer, 2013, pp. 176-186