# A Comparative Analysis of Machine Learning Algorithms for Customer Segmentation

Jaydeep Bohra, Ved Bansal
*Students of B. Tech Computer Science, Jaypee University of Engineering and Technology, Guna (MP), India*

**Abstract: This study investigates the effectiveness of various machine learning algorithms in the context of customer segmentation, a crucial aspect of Customer Relationship Management (CRM). With the increasing complexity of consumer behaviour and preferences, traditional segmentation methods are becoming less effective. We focus on four widely used algorithms: Logistic Regression, Decision Trees, Random Forests, and AdaBoost. By applying these algorithms to a retail customer dataset, we evaluate their performance based on accuracy, precision, and other relevant metrics. Our findings indicate that AdaBoost outperforms the other algorithms in terms of accuracy, while Logistic Regression demonstrates strong performance in scenarios with less complexity. This paper discusses the strengths and limitations of each algorithm, providing insights for researchers and practitioners in the field of customer analytics and marketing strategy.**

## INTRODUCTION

In today's competitive market landscape, understanding customer preferences has become paramount for businesses seeking to enhance customer satisfaction and drive revenue growth. Customer Relationship Management (CRM) practices have evolved to prioritize customer-centric strategies, with customer segmentation emerging as a cornerstone of effective marketing and service delivery [1]. By segmenting customers based on various attributes, companies can tailor their offerings to meet specific needs and preferences, thereby maximizing the effectiveness of their marketing efforts.

Traditional customer segmentation methods, which often rely on demographic, behavioural, and geographic factors, face significant limitations in exploring deeper insights into consumer behaviour. As the volume of data generated by consumers continues to grow, the need for advanced analytical techniques has become increasingly evident. Machine learning, with its ability to uncover complex patterns in large datasets, offers a promising solution to the challenges faced by traditional segmentation methods [5].

## 1. Methods

This study evaluates the performance of four machine learning algorithms on a dataset derived from retail transactions. The dataset, known as the "Customer Segmentation" dataset, was published by F. Daniel in 2015 and includes transaction information for approximately 4000 customers over a one-year period [7]. The algorithms under investigation include:

### 1.1 Logistic Regression

Logistic Regression is a statistical method used for binary classification that models the probability of a particular class based on one or more predictor variables. It employs a logistic function to map predicted values to probabilities [11]. While it is computationally efficient and interpretable, its linear nature limits its effectiveness in capturing complex relationships within the data.

### 1.2 Decision Trees

Decision Trees are a popular classification technique that recursively partitions the data into subsets based on the value of input features. The model is represented as a tree structure, where each internal node corresponds to a feature, each branch represents a decision rule, and each leaf node represents an outcome [13]. Decision Trees are intuitive and can model nonlinear relationships, but they are prone to overfitting, especially when the tree becomes too deep.

### 1.3 Random Forests

Random Forests build upon the Decision Tree algorithm by creating an ensemble of multiple trees to improve classification accuracy and robustness. Each tree is trained on a random subset of the data,

and predictions are made by aggregating the outputs of all trees [15]. This method reduces the risk of overfitting and enhances the model's ability to generalize to unseen data.

1.4 AdaBoost

Ada Boost, or Adaptive Boosting, is an ensemble learning technique that combines multiple weak classifiers to form a strong classifier. The algorithm focuses on instances that are misclassified by previous classifiers and adjusts their weights to improve accuracy [17]. This iterative approach allows AdaBoost to enhance the performance of weak learners, making it highly effective for classification tasks.

2. Data and Evaluation

The dataset utilized in this study includes transaction details such as invoice number, date, stock code, description, quantity, unit price, consumer ID, and country. This comprehensive dataset allows for a thorough analysis of customer behaviour, including purchase frequency and preferences.

To evaluate the performance of the algorithms, we employed several metrics, including:

- Accuracy: The proportion of correctly classified instances out of the total instances.
- Precision: The ratio of true positive predictions to the total predicted positives, providing insight into the model's ability to avoid false positives.
- Confusion Matrix: A tool that summarizes the performance of a classification algorithm by displaying the counts of true positive, true negative, false positive, and false negative predictions [8][9].

## RESULTS

The performance of each algorithm was assessed based on the aforementioned metrics. The results indicated that AdaBoost achieved the highest accuracy, followed by Logistic Regression, Random Forests, and Decision Trees. Specifically, AdaBoost's ability to focus on challenging instances significantly contributed to its superior performance.

The precision of Logistic Regression was found to be comparable to that of AdaBoost, particularly in scenarios where the dataset exhibited linear relationships. In contrast, Decision Trees demonstrated the lowest accuracy, primarily due to their sensitivity to noise and the potential for overfitting.

## DISCUSSION

The findings of this study highlight the strengths and limitations of each machine

## REFERENCES

[1] Hajiha A, Radfar R, Malayeri S. Data mining application for customer segmentation based on loyalty: An iranian food industry case study. 2011 IEEE International Conference on Industrial Engineering and Engineering Management. IEEE, 2011: 504 - 508.

[2] Das S, Nayak J. Customer segmentation via data mining techniques: state-of-the-art review. Computational Intelligence in Data Mining: Proceedings of ICCIDM 2021, 2022: 489 - 507.

[3] Online Retail. (2015). UCI Machine Learning Repository. https://doi.org/10.24432/C5BW33.

[4] Hasnain M, Pasha M F, Ghani I, et al. Evaluating trust prediction and confusion matrix measures for web services ranking. IEEE Access, 2020, 8: 90847 - 90861.

[5] Buya S, Tongkumchum P, Owusu B E. Modelling of land-use change in Thailand using binary logistic regression and multinomial logistic regression. Arabian Journal of Geosciences, 2020, 13: 1 - 12.

[6] Hu Z, Lo C P. Modeling urban growth in Atlanta using logistic regression. Computers, environment and urban systems, 2007, 31 (6): 667 - 688.

[7] Farid D M, Zhang L, Rahman C M, et al. Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert systems with applications, 2014, 41 (4): 1937 - 1946.

[8] Dimitriadis S I, Liparas D, Alzheimer's Disease Neuroimaging Initiative. How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer's disease: from Alzheimer's disease neuroimaging initiative (ADNI) database. Neural regeneration research, 2018, 13 (6): 962.

[9] Shahraki A, Abbasi M, Haugen Ø. Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost.

Engineering Applications of Artificial Intelligence, 2020, 94: 103770.