The Seybold Report

ISSN 1533-9211

DOI 10.17605/OSF.IO/8CKA3

TO BUY OR NOT TO BUY: COMPARISON OF MACHINE LEARNING TECHNIQUES TO PREDICT ONLINE SHOPPING PREFERENCE OF CUSTOMERS

DAKSH KAPOOR[1], ACHIRANGSHU CHAKRABORTY[2] and SUNITA DANIEL[3*]

[1, 2, 3] Lal Bahadur Shastri Institute of Management, Plot 11/7, Dwarka, Sector-11, New Delhi, India.
[*]Corresponding author Email: sunitadaniel@lbsim.ac.in

Abstract

In the last decade, and especially since 2020, online purchasing has become ubiquitous. Most customers prefer shopping online over visiting physical stores since it is more convenient and easier to shop from the comfort of their own home or workplace. On the other hand, the advantage of physical stores where products can be handled or tried on before purchasing is undeniable. Although it might be easy to determine the preferences and intentions of consumers who visit physical stores to make their purchases, it is harder to decipher the intentions and behavioral patterns of online shoppers, especially in large marketplaces that bring together a variety of products and sellers. This study has aimed to classify customers using machine learning techniques based on whether they complete a purchase using various browsing parameters and other dimensions. The analysis was carried out using secondary data obtained from Kaggle Machine Learning Repository. Bagging and boosting algorithms were used to predict purchasing intention of online shoppers. Since the dataset was highly unbalanced multiple techniques had to be used to balance them. It was found that the month of May had the highest revenue, and also the maximum number of customers making repeated visits to the website. Moreover, the month of May also had the maximum number of special days. The Gradient Boosting algorithm gave the highest accuracy in prediction of consumer behavior, however when Up sampling was performed, Light GBM gave the highest accuracy and for Down sampling, Random Forest gave the highest accuracy in prediction.

Keywords: Machine Learning, Exploratory Data Analysis, SMOTE, Near Miss Algorithm, Bagging, Boosting

1. INTRODUCTION

In the recent times with the onset of COVID-19 there has been a sudden rise in online shopping as the traditional shops were all closed and the only option available for customers was shopping on the internet. Due to closing of all shops even most businesses started setting up online shops and today almost every business has an online portal which allows customers to shop from the comfort of their homes. People merely need an internet connection and a digital payment to have their purchase delivered to them.

With the growing popularity of e-commerce, online shoppers look for various ways to understand the quality and other details products better. The various online reviews is a great source for shoppers to get a better understanding of the products. Online reviews, especially negatively distributed reviews play an important role in determining purchase intention [1]. The buying decision is also directly influenced by the buyer's behaviour. The stimulus-response model was the first approach used to explain consumer purchasing behavior [2]. It was noted that buyer consciousness is influenced by marketing and environmental cues. The S-O-R framework was effectively operationalized by various researchers in various settings to

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/8CKA3

explain consumer decision making process. The S-O-R Framework concluded that offline vendor ques have a strong and positive impact on the online purchasing intention of consumer. [2]

Since the advent of machine learning many researchers have used different algorithms to predict purchase intention. One such research concluded that the characteristics that influence attitudes toward online purchase are perceived utility, perceived simplicity of use, innovativeness, and perceived rewards [3]. The trust-antecedent perceived risk and the technology-antecedent perceived ease-of-use both have a direct impact on the attitude toward online shopping [4]. Using clickstream and session information data, accurate and scalable purchasing intention prediction for virtual shopping environments is also possible [5]. As technology advanced more real- time prediction models were being built to predict purchase intention. A real-time online consumer behavior prediction algorithm that anticipated if a visitor would purchase as soon as they arrive at the website used machine learning algorithm like random forest [6, 7]. One of the most well-known issues in the datasets, is the imbalance class categorization. Even though most of the customers log into a website with the intention of purchasing, a very small percentage of them end up purchasing. This is the main reason for the imbalanced dataset. Various techniques such as SMOTE and near Miss are available to balance the dataset [8, 9]. The bagging machine learning models such as Decision Tree, Random Forest and Neural Networks have also been used in predicting the customer purchase intention [5, 6, 9]. In this paper we use bagging and boosting machine learning techniques for prediction. The prediction is carried out for the original dataset, down sampled dataset using near-Miss and up sampled dataset using SMOTE technique.

2. THEORY AND FORMULA

Machine Learning algorithms are used to identify patterns in a dataset and test the accuracy of the predicted data. There are two major categories of machine learning algorithms which are bagging and boosting algorithms. Some of the bagging algorithms are Random Forest and boosting algorithms are AdaBoost, Gradient Boost and XG Boost algorithms. A Random Forest Classifier is a collection of several decision trees that work together to produce a better forecast. The primary variable in a random forest classifier is the number of estimators (the number of distinct decision trees to be considered for prediction). In a random forest approach, each decision tree makes a class prediction, and the class with the most votes in terms of target variable becomes the model's overall prediction. The AdaBoost algorithm creates a model and assigns equal weights to all data points. It then applies larger weights to incorrectly categorized points. In the following model, all points which are incorrectly categorized are given more weight. It will continue to train models until a smaller error is returned. Gradient Boosting is a prominent technique for boosting. In gradient boosting, each prediction corrects the inaccuracy of its previous stage. Unlike Adaboost, the weights of the training instances are not changed; instead, each predictor is trained using the predecessor's residual mistakes as labels. The XG Boost algorithm generates decision trees in a sequential fashion. All the independent variables are given weights, which are subsequently put into the decision tree, which predicts results. The weight of factors that the tree predicted incorrectly is raised, and these variables are

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/8CKA3

subsequently put into the second decision tree. Light GBM is a gradient boosting framework based on decision trees that improves model efficiency while reducing memory use. It employs two innovative approaches: Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB) that overcomes the constraints of the histogram-based approach employed in all GBDT (Gradient Boosting Decision Tree) frameworks. The performance of different models is evaluated using accuracy, a popular measure that refers to the system's ability to reliably predict the class label of new or unknown data. The following formula is used to calculate accuracy.

$$Accuracy = \frac{TP + TN}{Total\ sample}$$

True positives (TP) is the number of correctly anticipated positive.

True negatives (TN) Is the number of correctly predicted negative

False positive (FP) = number of observations projected as positive when they are negative.

False negative (FN) = number of observations projected as negative when, they are positive.

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate and the formula is given by:

$$Precision = \frac{TP}{TP + FP}$$

Recall is the ratio of correctly predicted positive observations to all observations in actual class – yes and is defined as

$$Recall = \frac{TP}{TP + FN}$$

F1 Score is another important metric of accuracy which is harmonic mean of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if there is uneven class distribution.

$$F1\ Score = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right)$$

When dealing with skewed database oversampling and under sampling techniques like SMOTE and near miss respectively are used to balance the dataset and the models are expected to perform better

3. EXPERIMENTAL SETUP

In this section, we have developed a framework to predict the purchase intent of online buyers based on various parameters such as bounce rate, administrative duration, information duration etc. The framework includes many machine learning techniques like Logistic Regression, Random Forest, AdaBoost, Decision Tree, XGBoost, Gradient Boosting and Light GBM to

ISSN 1533-9211

DOI 10.17605/OSF.IO/8CKA3

predict a customer's intent to purchase a product. These techniques were further compared to find the best technique based on the accuracy metrics. Figure 1 is a flowchart showing the various steps to be followed in analysing the dataset

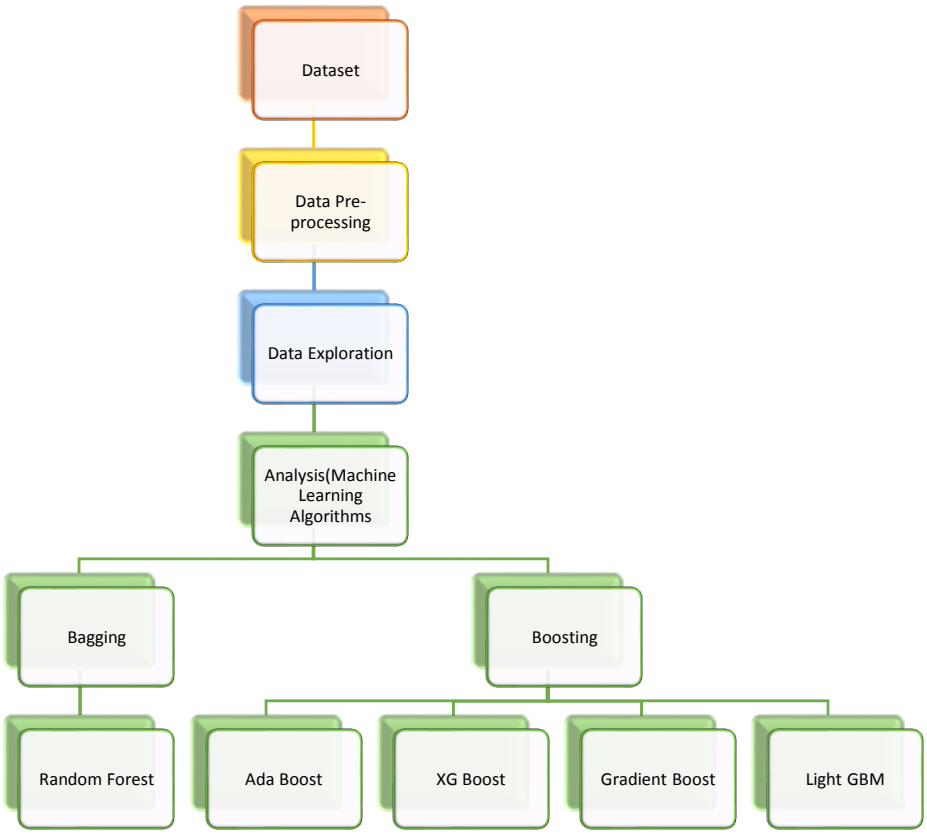

Figure 1: Flow chart showing the framework followed in analysis

Data Description

The dataset was taken from the Kaggle platform and it is about purchasing intentions of online customers. The main aim of the data is to classify customers in terms of whether they purchase something or not based on various browsing parameters and other dimensions that are categorised into ten numerical and eight categorical variables. This dataset was collected in real-time in 2018 and can be found in: https://www.kaggle.com/datasets/imakash3011/online-shoppers-purchasing-intention-dataset. The dataset contains around 12330 observations with 18 columns of different parameters. These parameters are explained in Table 1. The target variable is "Revenue" which tells us whether a customer has purchased of not.

It is observed that the data does not contain any missing values. The dataset is highly imbalanced in terms of revenue generation i.e 85% values are False while 15% of data set have

ISSN 1533-9211

DOI 10.17605/OSF.IO/8CKA3

revenue as True. The imbalance in the dataset can also be observed from Figure 2 that describes our target variable wherein it is observed that 10422 instances occur where the revenue is not generated for an ecommerce platform while 1908 instances are recorded in the data which depicts the revenue being generated for an ecommerce platform. Also, it has been found that parameters Administrative Duration, Informational Duration, ProductRelated_Duration, Bounce Rates, Exit Rates and Page Values are the only continuous variables and they are right skewed in nature.

Table 1: Description of the various parameters of the dataset

S.no	Feature Name	Description
1	Administrative	How many pages of this type the person visited.
2	Administrative Duration	Time spent on that page in seconds.
3	Informational	How many pages of this type the person visited.
4	Informational Duration	Time spent on that page in seconds.
5	Product Related	How many pages of this type the person visited.
6	ProductRelated_Duration	Time spent on that page in seconds.
7	Bounce Rates	The percentage of visitors who enter the website through that page and exit without triggering any additional tasks.
8	Exit Rates	The percentage of page views on the website that end at that specific page.
9	Page Values	The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction.
10	Special Day	The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother's Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction.
11	Month	The month of transaction.
12	Operating Systems	The operating system the user is using
13	Browser	The browser, the user is using
14	Region	The region ,the person is browsing from
15	Traffic Type	It is categorical and has been pre labeled, with 20 traffic types, the meaning is unclear.
16	Visitor Type	Tells the type of visitors, there can be 3, new, returning and other.
17	Weekend	False means, purchase was not on weekend and True if it is on weekend.
18	Revenue	False means, no purchase, True means a purchase occurred.

4. RESULTS AND DISCUSSIONS

The various results were obtained by Exploratory Data Analysis and application of the different machine learning algorithms using Python3.

ISSN 1533-9211

DOI 10.17605/OSF.IO/8CKA3

Figure 2: Bar chart showing the imbalance in target variable

Factors governing revenue generation:

As far as the revenue generated is concerned, from Table 2, the maximum revenue was generated in the month of November. November being the month generating maximum revenue may be because it is the beginning of the celebrations of Christmas and the dataset in consideration is from a western country. From Table 2, it can also be seen that maximum number of visitors to the ecommerce platform are in the month of May.

Table 2: Cross tabulation of Revenue generated in specific months

Month	Aug	Dec	Feb	Jul	June	Mar	May	Nov	Oct	Sep
Revenue										
FALSE	357	1511	181	366	259	1715	2999	2238	434	362
TRUE	76	216	3	66	29	192	365	760	115	86

Moreover, from Figure 2 it is observed that maximum revenue was generated during the weekdays. The main reason for this might be that most of the customers preferred going out on weekends for shopping and spend their time on the screen mostly during weekdays.

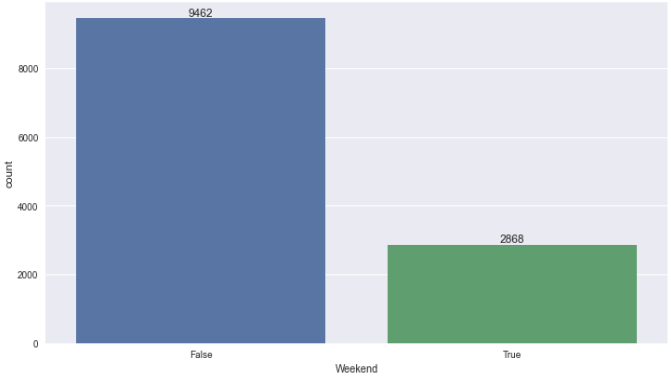

Figure 3: Revenue generated on weekends

ISSN 1533-9211

DOI 10.17605/OSF.IO/8CKA3

There are 3 types of visitors to the website namely new visitors, returning visitors and others (Table 1). New Visitors of the website contributed to 25% of the revenue and only 13% of the visitors returned to buy the product implying that most customers bought impulsively. This is shown in Table 3.

Table 3: Cross Tabulation showing the effect of Visitor Type on Revenue Generation

Count of Revenue	**Revenue**		
Visitor Type	**FALSE**	**TRUE**	**Grand Total**
New Visitor	1272	422	1694
Other	69	16	85
Returning Visitor	9081	1470	10551
(blank)			
Grand Total	**10422**	**1908**	**12330**

The important features for revenue generation can be obtained from the machine learning algorithm - Random Forest. Page Values, Month, Exit Rate, ProductRelated_Duration and Product Related are the top 5 important parameters, and this is depicted in Figure 3.

Relationship between the parameters of the dataset:

A correlation between the various parameters is also plotted in Figure 4 and it was found that the variables “Exit Rates” and “Bounce Rates” are strongly correlated with each other having correlation coefficient as 0.91. “Page Values” is the variable which is highly correlated with the target variable “Revenue”. Has the highest correlation.

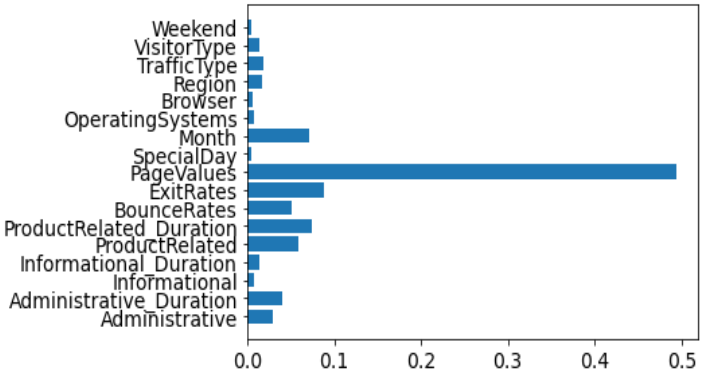

Figure 4: Feature importance of the dataset using Random Forest Algorithm

ISSN 1533-9211

DOI 10.17605/OSF.IO/8CKA3

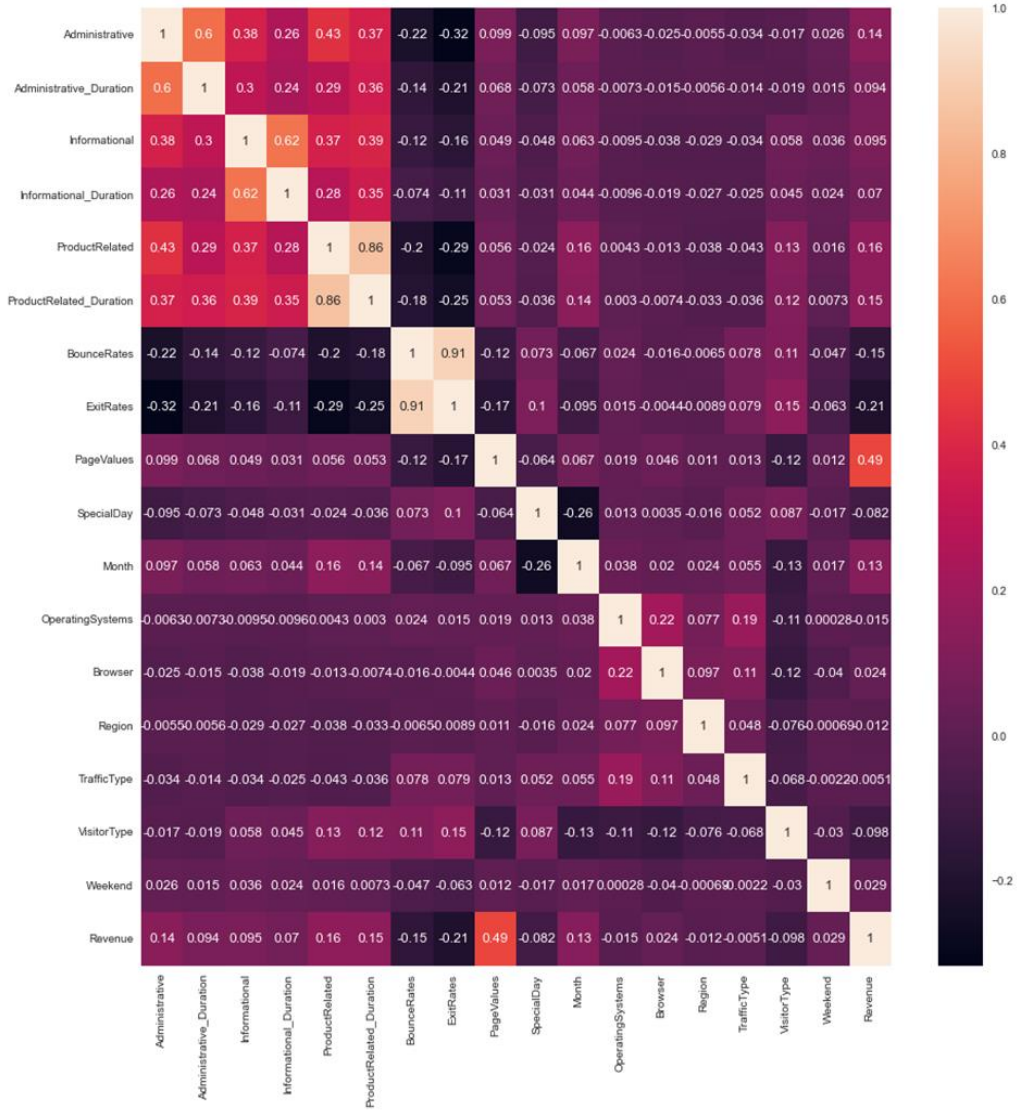

Figure 5: Correlation matrix of variables

Comparison of Machine Learning Algorithms:

Various machine learning algorithms were applied to the given dataset. The dataset was up sampled using SMOTE (Synthetic Minority Oversampling technique) and down sampled using the Near Miss Algorithm. The comparison of the results is shown in Table 4.

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/8CKA3

Table 4: Comparison of the accuracy of the bagging and boosting algorithms

Model	Normal sampling	With SMOTE	With Near miss
Random Forest	90.619	88.7537	88.7266
Adaboost	89.4295	87.3749	50.6353
Gradient Boost	90.8623	88.267	49.662
XG boost	89.8621	89.1051	50.1757
Light GBM	90.592	89.1862	50.6353

It is seen that for the normal or original dataset, Gradient Boost gives the highest accuracy and for the up sampled dataset, the Light GBM boosing algorithm gives the best prediction. The boosting algorithms give a very poor accuracy when the dataset is down sampled using the Near Miss algorithm.

Hence, the bagging model Random Forest gives a better accuracy for all types of datasets.

5. CONCLUSION

This study helps us to find the customer behavior pattern on ecommerce platforms and the various parameters which increases the revenue generation. The various parameters to enable a good revenue were identified using Exploratory Data Analysis. It is very important to know the amount of time that a customer spends on a particular webpage and how many webpages of the product the customer visits.

A comparison between the various machines learning algorithm was carried out for the original dataset. It was observed that the boosting algorithm fared badly when using the Near Miss Algorithm (down sampled dataset). The boosting algorithm gave a good accuracy for the original dataset as well as up sampled data. The Random Forest Algorithm performed equally well in all the three cases - original dataset, up sampled dataset and the down sampled dataset. For these types of datasets, the bagging algorithm gives a good accuracy. Hence Random Forest Algorithm can be applied to rightly predict whether a customer would generate revenue or not.

REFERENCES

1. Yang, Jing, Rathindra Sarathy, and JinKyu Lee. (2016) "The effect of product review balance and volume on online Shoppers' risk perception and purchase intention." Decision Support Systems 89: 66-76.
2. Kaur, Sukhwinder, Amit Kumar Lal, and Sarbjit Singh Bedi. (2017) "Do vendor cues influence purchase intention of online shoppers? An empirical study using SOR framework." Journal of Internet Commerce 16, no. 4 : 343-363.
3. Mostafa, Rania B., and Hassan Naim Hannouf. (2022) "Determinants of Online Purchase Intention of Apparel Products in Lebanon." International Journal of Online Marketing (IJOM) 12, no. 1 :1-18.
4. Van der Heijden, Hans, Tibert Verhagen, and Marcel Creemers. (2003) "Understanding online purchase intentions: contributions from technology and trust perspectives." European journal of information systems 12, no. 1: 41-48.
5. Sakar, C. Okan, S. Olcay Polat, Mete Katircioglu, and Yomi Kastro. (2019) "Real-time prediction of online

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/8CKA3

shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks." Neural Computing and Applications 31, no. 10: 6893-6908.

6. Baati, Karim, and Mouad Mohsil. "Real-time prediction of online shoppers' purchasing intention using random forest." In IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 43-51. Springer, Cham, 2020.
7. Esmeli, Ramazan, Mohamed Bader-El-Den, and Hassana Abdullahi. (2021) "Towards early purchase intention prediction in online session based retailing systems." Electronic Markets 31, no. 3: 697-715.
8. Kurniawan, I., M. F. Akbar, D. F. Saepudin, M. S. Azis, and M. Tabrani. "Improving The Effectiveness of Classification Using The Data Level Approach and Feature Selection Techniques in Online Shoppers Purchasing Intention Prediction." In Journal of Physics: Conference Series, vol. 1641, no. 1, p. 012083. IOP Publishing, 2020.
9. Li, Na, Chongyi Gong, and Dongqin Lv. "(2022) Real-Time Prediction of Cross-Border e-Commerce Spike Performance Based on Neural Network and Decision Tree." Wireless Communications and Mobile Computing 2022.