Prediction of the Purchase Intention of Users on E-Commerce Platforms using Gradient Boosting

Yannick Kiki, Vinasetan Ratheil Houndji

Abstract*: In this paper, we propose a system that is able to forecast the purchase intention of users visiting e-commerce platforms from data collected as they browse on these websites. We use the Online Shoppers Purchasing Intention Dataset available at the University of California Irvine Machine Learning Repository. Thanks to some feature engineering methods, we deeply study the correlation between the various information. We also derive new information / features from the dataset by inference. The most relevant data is fed to gradient boosting, artificial neural networks and other algorithms in order to forecast whether or not a user intends to make a purchase. We evaluate the performances with the precision metric and the F1-Score. The experiments show that our gradient boosting model performs better than the state-of-the-art models thanks to the new features used. This also confirms that, in addition to being interpretable, some classic machine learning models such as gradient boosting can be very competitive compared to neural networks. This system thus conceived can allow e-commerce platforms to identify users intending to make a purchase. This gives them the possibility of offering personalized solutions to their potential customers in order to better attract them and guarantee their purchase, which will imply increased sales and better customer satisfaction.

***Keywords*: e-commerce, feature engineering, gradient boosting, machine learning.**

I. INTRODUCTION

The development of computer networks and the Internet has revolutionized ways of doing things in several business sectors including commerce. Today, the reliability and success of e-commerce are no longer to be proven. Thanks to the multiple advantages that this system offers (such as increased visibility, disappearance of borders, reduced costs, etc.), it turns out to be a very good option for any merchant wishing to make the most of his activity. However, classic e-commerce has some drawbacks. For example, customers who intend to make a purchase may leave an e-commerce site without having made that purchase due to a lack of support. This situation is generally not encountered in the context of physical commerce because there are salespeople who closely follow the actions of the customer and could then make proposals and recommendations to them. At a time when artificial intelligence has many applications and encouraging results, it would be interesting to analyze its usefulness in correcting this issue. We are interested in this work in the application of machine learning techniques in the design of a system for detecting the purchase intention of users of e-commerce platforms.

Revised Manuscript Received on October 28, 2020.

Yannick KIKI, Student, Department of Computer Engineering, Ecole Polytechnique d'Abomey-Calavi, University of Abomey-Calavi, Godomey, Benin.

Vinasetan Ratheil HOUNDJI, Department of Computer Engineering, Ecole Polytechnique d'Abomey-Calavi, University of Abomey-Calavi, Godomey, Benin.

This system will allow better assistance of potential customers and then an increase of income. Our work relies on [8]. The authors propose a real-time online shopper behavior analysis system based on a multi-layer perceptron model which predicts the purchasing intention of the visitor using aggregated pageview data kept track during the visit along with some sessions and user information. In our work, we consider the same dataset but, instead of directly providing all the features of the dataset to the various models, we make a preliminary statistical study in order to derive new features from those existing. Then, we determine the most relevant features in the new set of features and thus make the training of the model more efficient. We also tried several model structures not tested in the previous work. The experiments show that with a classic gradient boosting algorithm, one could obtain very good results. The paper proceeds as follows. Next section provides an overview on the state-of-the-art. Section III gives some details about the dataset used. Then the methodology is described in Section IV. Finally, Section V shows the main results obtained.

II. REVIEW OF LITERATURE

Several works have been done in order to improve e-commerce and to analyze users' behaviour. Let us see some examples. Carmona et al. [1] present web usage mining as a solution to allow retailers to have a more successful business. Their work focused on a presentation of the components of web mining as well as on the main data preprocessing tasks for web usage mining. Rajamma et al. [2] examine consumer behavior at the final stages of transaction culmination and find out that perceived transaction inconvenience is the major predictor of shopping cart abandonment. Ding et al. [3] show an individual-level, dynamic model with concurrent optimal page adaptation that learns users' real-time, unobserved intent from their online cart choices, and immediately perform optimal Web page adaptation to enhance the conversion of users into buyers. Albert et al. [4], developed a model to guide the design and the continuous management of customer-centric web-based systems, such as e-commerce web sites and the case study showed considerable measured improvement in the effectiveness of these websites. Awad et al. [5],
proposed a new modified Markov model to predict the next set of Web pages that a user may visit based on the knowledge of the previously visited pages. Budnikas [6], used a Naïve Bayes classifier and a feedforward neural network for the classification of online patterns of user behaviour as well as for an estimation of a website component that has the highest impact on a fulfilment of business objectives by a user. Fernandes et al.

[7] present the use of data mining techniques to analyze sequences of accessed pages by the users in order to understand customer purchasing engagement and real-time purchase likelihood. Sakar et al. [8] used multilayer perceptron and LSTM recurrent neural networks to propose a real-time online shopper behavior analysis system which simultaneously predicts the visitor's shopping intent and Web site abandonment likelihood. Our work relies on this latter.

III. DATASET

In this work, we use the Online Shoppers Purchasing Intention Dataset available on UCI Machine Learning Repository [9]. It is the same dataset used in [8] and we aim to use some new methodologies to better take advantage of it. This dataset comes from the collection of data on the behavior of users of e-commerce platforms during their visit to the said platform. It consists of vectors of features belonging to 12330 sessions. The dataset was formed so that each session would belong to a specific user over a period of one year to avoid any tendency to a specific campaign, a special day, a user profile, a period. This dataset consists of 10 numerical and 8 categorical attributes. These attributes are shown in Table- I and Table- II, respectively. Of the 12,330 sessions of the dataset, 84.5% (10,422) were negative class samples (did not end with shopping), and the rest (1908) were positive class samples (ending with purchases).

The attribute that indicates whether the purchase has been made or not is named "Revenue" and will be used as the class label. This attribute consists of boolean values. "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another. The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site. The value of "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session. The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session. The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction. 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 a transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentine's day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8. The dataset also includes operating system, browser, region, traffic type, visitor type as 'returning' or 'new visitor', a boolean value indicating whether the date of the visit is weekend, and month of the year [8].

Table- I: Numerical features

Feature	Min	Max	Std
Administrative	0	27	3.32
Administrative Duration	0	3398	176.70
Informational	0	24	1.26
Informational Duration	0	2549	140.64
Product Related	0	705	44.45
Product Related Duration	0	63973	1912.25
Bounce Rate	0	0.2	0.04
Exit Rate	0	0.2	0.05
Special Day	0	361	18.55
Page Value	0	1.0	0.19

Table- II: Categorical features

Feature	Number of values
Operating System	8
Browser	13
Region	9
Traffic Type	20
Visitor Type	3
Weekend	2
Month	12
Revenue	2

IV. METHODOLOGY

Our aim is to predict the *Revenue* attribute of our dataset. To have the best possible prediction system using the data at our disposal, we have followed these steps. In a first step named feature engineering, we carried out a statistical study on the various parameters and their influence on the purchase intention of a user. In a second step named models selection, we have deeply studied the various algorithms used in the context of data prediction more particularly in a problem of classification and tried to identify those that performs best on our problem with models training and evaluation steps. We describe each of these steps below.

A. Feature Engineering

In our dataset, the various information are represented by the features listed in Table- I and Table- II. We studied these features and added new ones using inference rules. It allows us to bring out of the dataset some information that existed but not explicitly.

Afterwards, we determined which features are the most interesting in our dataset using several techniques such as the chi-square test, the extra trees classification, the recursive feature elimination algorithm and pearson correlation.

B. Models Selection

After the statistical study of the data, our aim is to select the most convenient model algorithm to use to implement our system. There are several models that theoretically promise high performance.

For this problem, we worked with models known to have good results in classification tasks. Some are classic machine learning models and others are neural network models: Gradient Boosting, Random Forest, Support Vector Machines, Artificial Neural Networks, Multi Layer Perceptron.

C. Models Training

During this phase, we train the various algorithm models selected with our dataset by considering various sets of features. Thanks to the phase of statistical study of the data in the feature engineering step (Section IV-A), we can know which features are more relevant but according to the technique used, the features being designated as the best were not always the same. The issue that follows is that we have to determine the number of features to use (among our features already ranked in order of importance in the previous step) and the algorithm model to use to have the most efficient final result.

The solution chosen is to test the different sets of features given by the statistical study by keeping the order of importance of the parameters and by considering that we could have the optimal set of features with the *k* best parameters. We also made mixtures of the sets of *x* best features obtained with the various methods and generated new sets of features that it would be interesting to test. We then passed each set of features on each algorithm model selected.

We used 80% of the data for training and 20% of the data to check whether the model derived from the training is actually able to predict results on data it has never encountered before.

D. Evaluation

After training the models on 80% of the data, we test them on the remaining 20% of the data. To be able to evaluate the performance of the results, we use several metrics namely:

True Positive Rate: Also called sensitivity. It is calculated as follows:

$$TPR = TP/(TP+FN)$$

with

TP: *True Positive*, **TN**: *True Negative*, **FP**: *False Positive*, **FN**: *False Negative.*

TPR is the probability that a result of the positive class is correctly recognized as being of the positive class during the test. In our case, this comes down to the probability that a customer intending to make a purchase is correctly recognized as so in the test.

True Negative Rate: Also called specificity. It is denoted *TNR* and is calculated as follows:

$$TNR = TN/(TN + FP)$$

TNR is the probability that a result of the negative class is correctly recognized as being of the negative class during the test. In our case, this comes down to the probability that a customer not intending to make a purchase is correctly recognized as so in the test.

Precision:

The precision makes it possible to know what proportion of identifications was actually correct.

$$Precision = TP/(TP + FP)$$

F1-Score:

In the statistical analysis of binary classification, the F1-Score is a measure of the accuracy of a test. It considers both the precision and the recall.

$$F1\text{-}Score = 1/(1/precision + 1/recall)$$

The recall corresponds to the number of items identified as positive among the positive class items. It is therefore equivalent to *TPR*.

The common point of all these metrics is that they increase proportionally with the number of correct predictions and conversely proportionally with the number of incorrect predictions.

V. RESULTS AND DISCUSSION

A. Feature Engineering

As mentioned above, the dataset used consists of 12,330 records of online user activity. Each record is described by taking various information through the 18 features available in the dataset. The first step in our statistical study is to determine what additional information we could infer from the information we have.

This is how we added the following parameters:

- **nPages Visited**: Total number of pages visited by the user on the store;
- **total Duration**: Total time in seconds spent by the visitor on the store;
- **avg Administrative Duration**: Average time spent by the visitor on pages linked to account management;
- **avg Informational Duration**: Average time spent by the visitor on information pages
- **avg Product Related Duration**: Average time spent by the visitor on the pages linked to the products;
- **avg Duration Per Page**: Average time spent by the visitor on a page of the store;
- **is First Semester**: Boolean value indicating if we are in the first half of the year;
- **is Jan, is Feb, is Mar, is Apr, is May, is Jun, is Jul, is Aug, is Sep, is Oct, is Nov, is Dec**: Boolean values indicating if we are in a giving month.

After generating these features, our dataset contains now 37 features. Then we evaluate the influence of our new set of features on our result class label. Table III shows the results of these tests.

Table- III: Top 10 most influential features

Rank	Chi-square test	Extra trees classification	RFE	Pearson
1	totalDuration	Administrative	InformationalDuration	PageValue
2	ProductRelatedDuration	AdministrativeDuration	totalDuration	ExitRate
3	PageValues	Informational	ProductRelatedDuration	nPagesVisited
4	Administrative Duration	InformationalDuration	isMay	ProductRelated
5	Informational Duration	ProductRelated	TrafficType	TotalDuration
6	nPagesVisited	ProductRelatedDuration	avgInformationalDuration	isNov

7	ProductRelated	BounceRate	AdministrativeDuration	ProductRelatedDuration
8	avgInformationalDuration	ExitRate	avgAdministrativeDuration	BounceRate
9	avgAdministrativeDuration	PageValue	avgDurationPerPage	Administrative
10	ProductRelatedDuration	SpecialDay	avgProductRelatedDuration	isFirstSemester

B. Training results

At this stage of the study, we know the most important features according to various tests. The problem that arises is to know what is the set of features to retain for our study. To maximize our chances of retaining the best set of features, we select several and test them all. The selection of the different set of features to be tested is made as follows:

- each set of n best features according to each method presented in Table- III and this for n ranging from 5 to 25;
- each set of features consisting of the mixture of the n best features of each method for n ranging from 2 to 11.

We obtain *95* sets of features to test. We then take each of the models selected, instantiate and train them independently on each set of features. The best results are presented in Table- IV. In this table, FS means Feature Set and ANN is Artificial Neural Network.

Table- IV: Best results obtained

FS	Model used	TPR	TNR	Precision	F1-Score
A	Gradient Boosting	**0.844**	**0.911**	**0.901**	**0.907**
B	ANN	0.839	0.909	0.899	0.905

with

- FS-A: Informational Duration, total Duration, Administrative Duration, Product Related Duration, Bounce Rates, Traffic Type, nPages Visited, avg Informational Duration, Exit Rates, Informational, Administrative, Product Related, is May, is Nov, Page Values.
- FS-B: Administrative, Administrative Duration, Informational, Informational Duration, Product Related, Product Related Duration, Bounce Rate, Exit Rates, Page Values, Special Day, Operating Systems, Browser, Region, Traffic Type.

C. Discussion

Table- V shows the best state-of-the-art results obtained on the same dataset and our gradient boosting model.

Table-V: Best results previous work on the same dataset

Model used	TPR	TNR	Precision	F1-Score
MLP	*0.84*	***0.92***	*0.87*	*0.86*
C4.5	0.79	0.85	0.823	0.82
SVM	0.75	0.94	0.85	0.82
Our Gradient Boosting model	**0.84**	0.91	**0.90**	**0.91**

We can notice that our performances are globally better than the ones obtained from the state-of-the-art. The Gradient Boosting model is highly competitive w.r.t. the other models. This confirms that it remains interesting to take into account classic machine learning models instead of systematically turning to deep learning for classification problems.

VI. CONCLUSION

In this article, we have proposed a system based on the gradient boosting algorithm which is able to predict the purchase intention of users of e-commerce platforms from data collected as they browse on these platforms. To achieve that, we have used the Online Shoppers Purchasing Intention Dataset and applied a strict methodology in order to make the most of this dataset as well as the most recent techniques and algorithms used in classification problems. This methodology consisted, in a first step, to make feature engineering on the dataset in order to infer new relevant information, that is, those that have the most influence on our result, the Revenue attribute, which describes if the user has made a purchase at the end of his navigation. Once the most relevant information had been identified, we looked at the machine learning algorithms that will allow us to make the most of this information. This is how in our study we worked with different machine learning algorithms, some being classic machine learning algorithms and others being artificial neural networks. We trained different prediction models with these algorithms and then we evaluated them using metrics such as TPR, TNR and F1-Score. This methodology allowed us to obtain interesting results with a model which performs better than the state-of-the-art models.

Model used	TPR	TNR	Precision	F1-Score
Gradient Boosting	**0.84**	0.91	**0.90**	**0.91**

It should also be noted that our best result was obtained with a gradient boosting model while we also tested with artificial neural network models. We can then point out the fact that classic machine learning models can perform better than deep learning models on classification problems. The system we offer enables online merchants to identify users on their platform who intend to make a purchase. This system can then be used to offer personalized solutions to these potential customers or to be able to contact them if they do not end up making a purchase and this ultimately implies increased income and better customer satisfaction.

REFERENCES

1. Carmona CJ, Ramı´rez-Gallego S, Torres F, Bernal E, del Jesús MJ, Garcıa S (2012) Web usage mining to improve the design of an e-commerce website: OrOliveSur. com. Expert Syst Appl 39(12):11243–11249
2. Rajamma, Rajasree K.; Paswan, Audhesh K.; and Hossain, Muhammad M., "Why do shoppers abandon shopping cart? Perceived waiting time, risk, and transaction inconvenience" (2009). *Business Faculty Publications*. 205.
3. Ding AW, Li S, Chatterjee P (2015) Learning user real-time intent for optimal dynamic web page transformation. Inf Syst Res 26(2):339–359

4. Albert TC, Goes PB, Gupta A (2004) A model for design and management of content and interactivity of customer-centric websites. MIS Q 28(2):161–182
5. Awad MA, Khalil I (2012) Prediction of user's web-browsing behavior: application of markov model. IEEE Trans Syst Man Cybern B Cybern 42(4):1131–1142
6. Budnikas G (2015) Computerised recommendations on e-trans- action finalisation by means of machine learning. Stat Transit New Ser 16(2):309–322
7. Fernandes RF, Teixeira CM (2015) Using clickstream data to analyze online purchase intentions. Master's thesis, University of Porto
8. Sakar, C.O., Polat, S.O., Katircioglu, M. *et al.* Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Comput&Applic* 31, 6893–6908 (2019). https://doi.org/10.1007/s00521-018-3523-0
9. UCI Machine Learning Repository. Available :https://archive.ics.uci.edu/. Last accessed: 19/08/2020.

AUTHORS PROFILE

Yannick KIKI is a computer engineering student at Ecole Polytechnique d'Abomey-Calavi in the University of Abomey-Calavi. He studied for two years in the Industrial Sector before specializing in the 3rd year in Computer Engineering and Telecommunications. In his last year of study, he focused on Computer Science and is currently preparing his defense for the engineering degree. His areas of interest are Machine Learning, Data Science, Software Engineering and Optimisation. During his student journey, he won the MIFY Artificial Intelligence Contest 2017 with the OBF team and was first-ranked in Benin for the Google Hash Code 2019 and 2020 with the JHYL team.

Vinasetan Ratheil HOUNDJI received a Ph.D. in Computer Science (from Universitécatholique de Louvain - UCL, Belgium &Universitéd'Abomey-Calavi - UAC, Benin) in 2017 after obtaining a Master of Science degree in Computer Science (from Ecole Polytechnique de Louvain, UCL, Belgium) in 2013 and an Engineer degree in Computer Science and Telecommunications (from Ecole Polytechnique d'Abomey-Calavi, UAC) in 2011. He co-founded the companyMachine Intelligence For You (MIFY) in 2017 and spent one year as Chief Executive Officer of this company. African Scientific Institute (ASI) Fellow since February 2019, he currently is lecturer-researcher atUAC, mainly in Artificial Intelligence and Operations Research/Optimization; and Chair of theAssociation for the Advancement of Artificial Intelligence (AAAI) Benin chapter.
His research interests are focused on Artificial Intelligence, Machine Learning, Constraint Programming, and Optimisation. https//ratheil.info