Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest

Karim Baati[1(✉)] and Mouad Mohsil[2]

[1] Teolia Consulting, 12–14 Rond-Point des Champs Elysées, 75008 Paris, France
[2] Teolia D.A., 12–14 Rond-Point des Champs Elysées, 75008 Paris, France
{karim.baati,mouad.mohsil}@teolia.fr

Abstract. In this paper, we suggest a real-time online shopper behavior prediction system which predicts the visitor's shopping intent as soon as the website is visited. To do that, we rely on session and visitor information and we investigate naïve Bayes classifier, C4.5 decision tree and random forest. Furthermore, we use oversampling to improve the performance and the scalability of each classifier. The results show that random forest produces significantly higher accuracy and F1 Score than the compared techniques.

Keywords: Real-time online shopper behavior · Marketing offers in online stores · Random forest

1 Introduction

Nowadays, a large majority of businesses are supported or carried out online. In order to foster the generated virtual environments, marketing offers stand for one of the most valuable strategies which can be employed. Historically, these offers were indiscriminately suggested to the whole visitors of a given e-commerce website. Afterward, being aware about the necessity to orient their marketing actions to the right target, online stores opted for a near-real time analysis of visitors' information. The purpose is to contact the most relevant users (for instance by phone or e-mail) in order to suggest offers which are likely to induce them to go back to the website and achieve an effective purchase.

Recently, a new trend has emerged among virtual shopping environments so that potential visitors are identified at the time they are browsing the website. By contrast to the near-real time model, the advantage behind that is to avoid the high risk of losing users once disconnected from the online store. Indeed, in such a model, we imitate an experienced salesperson who struggles to retain potential visitors by providing a range of customized marketing actions which are likely to encourage them to buy. The latest study suggesting such a strategy in e-commerce websites could be found in [1] where authors proposed a system with two modules which predicts the purchasing intent of the visitor by using some

Published by Springer Nature Switzerland AG 2020
I. Maglogiannis et al. (Eds.): AIAI 2020, IFIP AICT 583, pp. 43–51, 2020.
https://doi.org/10.1007/978-3-030-49161-1_4

session and user information along with aggregated pageview data kept track during the visit. The first module of this system is used to determine whether the user should be offered content and the second module is triggered only if the user is likely to abandon the site. Though the fact that the system proposed in [1] is appealing in terms of efficiency and scalability, the risk of abandon it implies stands for a problem which is not to be neglected.

In this paper, we do not aim to propose a new model that substitutes systems like the one of Sakar et al. [1]. On the contrary, our objective is to consolidate such systems by trying to retain the maximum number of potential visitors. In this scope, we suggest a system that allows to detect users with high purchasing intention as soon as they connect to an e-commerce website. For this purpose, we rest on the same data used in [1] but we only keep those pertaining to session and user information. As that will be detailed in this paper, by establishing our system, we aspire to be part of a global system that starts by proposing a first type of marketing offers to potential visitors once connected to the website. Later, that system calls a second subsystem (like the one of Sakar et al. [1]) to suggest more generous offers to visitors who did not carry out an effective purchase at the first stage but who displayed a high purchasing intention after a certain clickstream.

The remainder of the paper is structured as follows. Related work is reported in Sect. 2. Next, Sect. 3 details different functional and technical aspects of our proposed system. Afterward, experimentation results are communicated in Sect. 4. Lastly, Sect. 5 concludes the paper and suggests some directions for future research.

2 Literature Review

Literature encompasses many studies which are turned towards categorization of online visits in e-commerce websites.

A first study of Mobasher et al. [2] assessed two different clustering techniques based on user transactions and pageviews in order to find out useful aggregate profiles that can be used by recommendation systems to achieve effective personalization at early stages of user's visits in an online store.

Later, the study of Moe [3] prepared the ground for a system which can take customized actions according to the category of a visit. For this reason, the author proposed a system which makes use of page-to-page clickstream data from a given online store in order to categorize visits as a buying, browsing, searching, or knowledge-building visit. The proposed system rests on observed in-store navigational patterns (including the general content of the pages viewed) and a k-means algorithm for clustering.

In [4], Poggi et al. proposed a system which handles the loss of throughput in Web servers due to overloading, by assigning priorities to sessions on an e-commerce website according to the revenue that will generate. Data were formed of clickstream and session information and Markov chains, logistic linear regression, decision trees and naïve Bayes were investigated in order to measure the probability of users' purchasing intention [4].

In [5] and [6], authors designed the prediction of purchasing intention problem as a supervised learning problem and historical data collected from an online bookstore were used to categorize the user sessions as browsing and buyer sessions. In this scope, Support vector machines (SVMs) with different kernel types and k-Nearest Neighbor (k-NN) were respectively investigated in [5] and [6] to carry out classification.

In a more recent study [7], Suchacka and Chodak constructed a new approach to analyze historical data obtained from a real online bookstore. The proposed approach is based on association rule discovery in customer sessions and aims to evaluate the purchase probability in an online session.

In [8], the author proposed a system that identifies the website component that has the highest business impact on visitors. To build such a system, a data set based on the Google Analytics tracking code [9] has been created. Moreover, naïve Bayes and multilayer perceptron classifiers have been explored for classification.

In [1], authors set up a real-time user behavior analysis system for virtual shopping environment which is made up of two modules. In the first module, the purchasing intention of the visitor is predicted using aggregated pageview data kept track during the visit along with some session and user information. Further, oversampling and feature selection preprocessing algorithms were applied to improve the effectiveness and the scalability of a set of supervised machine learning techniques. The highest accuracy of 87.24% and F1 Score of 0.86 were obtained with a Multilayer Perceptron Network (MLP). In the second module, authors used a Long Short-Term Memory-based Recurrent Neural Network (LSTM-RNN) based on sequential clickstream data to produce the probability estimate of visitor's intention to leave the site without completing the transaction. Within the scope of the entire system proposed in [1], the first module is triggered only if the second module generates a greater value than a predetermined threshold and the final objective consists in deciding whether to offer a content to the online visitor.

3 Proposed System

In this section, we describe the functional aspects related to our proposed system. Next, we depict the dataset used for validation as well as the machine learning techniques which have been explored for classification.

3.1 Functional Description

Based on previous studies, we can notice that no work has addressed the purchasing intention of online visitors immediately after they connect to the website. By contrast to these studies, our system aims to detect users with high purchasing intention as soon as they connect to the e-commerce website and to offer content only to those who intend to complete a transaction. The advantage behind that is to avoid the risk of losing potential visitors who sometimes disconnect for

trivial reasons (arrival of a guest at home, reception of a phone call, etc.). That risk should not be neglected since it naturally impacts the effectiveness of any system with a business objective (for instance each of the systems introduced respectively in [1,5,6] and [7]).

As mentioned earlier, through our proposed system, we do not aspire to replace the previous systems that have tackled the same problem by using clickstream data. On the contrary, we aim to reinforce that systems by trying to interest the maximum number of potential users. Indeed, our model could be used as a part of a more global structure as shown in Fig. 1. This global structure starts by proposing a first type of marketing offers to potential visitors once connected to the website and appeals later a second subsystem (like the one proposed in [1]) to suggest more generous offers to visitors who did not accomplish an effective purchase at the first stage but who showcased a high purchasing intention after a certain clickstream (Fig. 1). In our opinion, such a global system can be investigated for e-commerce websites as a new tool that could be useful in terms of effective use of time, purchase conversion rates and sales figures.

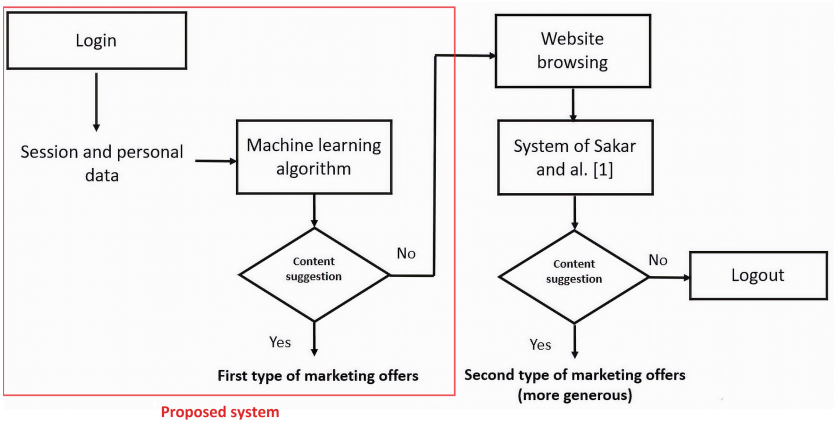

Fig. 1. Flowchart of the proposed system and its positioning within a more global desired system

3.2 Data Description

As mentioned earlier, we reposed on the same dataset used in [1] but we only keep the part pertaining to session and user information. Features related to the selected data are depicted in Table 1. As reported in [1], features belong to 12330 sessions and the data were constituted so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. Moreover, among the 12330 sessions

in the dataset, 84.5% (10422) were negative class samples that did not finalize a transaction and the rest (1908) were positive class samples ending with purchasing [1].

Based in Table 1, we can notice that, except the "Day" feature, all the rest of features are categorical. In order to homogenize the features types, we used a binning process in which the "Day" variable is discretized to 5 discrete levels [10].

Table 1. Features used for system validation

Feature	Feature description	Feature type
Day	Closeness of the site visiting time to a special day	Numerical
Operating Systems	Operating system of the visitor	Categorical
Browser	Browser of the visitor	Categorical
Region	Geographic region from which the session has been started by the visitor	Categorical
Traffic	Traffic source by which the visitor has arrived at the website	Categorical
Visitor	Visitor type as "New Visitor", "Returning Visitor" and "Other"	Categorical
Weekend	Boolean value indicating whether the date of the visit is weekend	Categorical
Month	Month value of the visit date	Categorical
Revenue	Class label indicating whether the visit has been finalized with a transaction	Categorical

3.3 Prediction

Technically, a challenging issue for the proposed system consists in finding out the most suitable machine learning technique for the prediction problem. In this context, specifications of input data are among main criteria which could be used to elect such a technique [11,12]. In our case, since features are categorical [13], three appropriate techniques for this type of features are selected, namely: naïve Bayes classifier, C4.5 classifier and random forest. In the following, fundamentals of each one of these techniques are briefly detailed.

- **Naïve Bayes classifier**
  Naïve Bayesian Classifier (NBC) is a probabilistic classifier which is based on Bayes fusion rule. It assumes the independence of the input features. That means that it considers each of these features to contribute independently to the probability of a class regardless of any possible correlations between

variables [14,15]. The final decision is assigned to the class with the highest probability. Despite its simplicity, NBC can often outperform more sophisticated classification methods [16].

- **C4.5 classifier**
  C4.5 is an extension of the earlier ID3 algorithm and stands for an algorithm used to generate a decision tree [17] which can be used for classification.
  At each node of the tree, C4.5 selects the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurses on the partitioned sublists.
- **Random forest classifier**
  Random Forest (RF) [18] is a well-known decision tree ensemble that is commonly used in classification. In many cases, RF showed a good performance that outstrips that of many other classification algorithms [19].
  As a decision tree ensemble, RF needs to construct several different decision trees. In order to accomplish that, each tree is set up considering a bootstrap sample set of the original training data. That consists in creating a new set sampling with replacement instances from the original set until getting the size of that original training data. Using one of the bootstrap sample sets of the training data, a random tree is obtained. In order to favor the diversity of the ensemble, each random tree follows a traditional top-down induction procedure with several modifications. At each step, when the best attribute is chosen, only a small subset of attributes from the dataset is considered. Considering only the subset of attributes chosen, we then compute the best attribute as in Classification And Regression Trees (CART) [20]. Each tree is built to its maximum depth and no pruning procedure is applied after the tree has been fully built. Lastly, to compute the predicted class for a sample, the predictions of the ensemble of decision trees are aggregated through majority voting.

4 Experiments and Results

To conduct experiments, data described in Sect. 3.2 are fed to naïve Bayes, C4.5 decision tree and random forest classifiers using 70% of dataset for training and the rest for validation [21,22]. Moreover, to ensure statistical significance, this procedure is repeated 100 times with random training/validation partitions.

Table 2 presents the average accuracy, sensitivity (true positive rate), specificity (true negative rate) and F1 Score for each classifier.

The results show that naïve Bayes classifier yields the highest accuracy rate on test data. However, a class imbalance problem emerges [23] since, for each classifier, the sensitivity is much lower than specificity (sensitivity is even null for naïve Bayes classifier). Indeed, it is clear that each classification technique tends to label the test samples as the majority class (negative class). This class

imbalance problem is a natural situation for the addressed problem since most of the online visits do not end with purchasing [24]. However, this issue should be properly handled since correctly identifying directed buying visits which are represented with positive class in the dataset is as crucial as identifying negative class samples.

To deal with class imbalance problem, we considered the use of the widely used technique to balance the training set before the learning phase, namely the "Synthetic Minority Oversampling Technique" (SMOTE) methodology [25]. To do that, 30% of the data set consisting of 12330 samples is first excluded for testing and the oversampling SMOTE method is applied to the remaining 70% of the samples.

The results obtained on the balanced dataset are reported in Table 3. As it is seen, the highest accuracy of 86.78% and F1 Score of 0.60 is obtained with the random forest classifier.

Table 2. Results obtained on class imbalanced dataset

Classifier	Accuracy (%)	Sensitivity	Specificity	F1 score
Naïve Bayes	90.04	0.00	1.00	0.00
C4.5	86.27	0.11	0.94	0.13
Random forest	83.64	0.09	0.91	0.10

Table 3. Results obtained with oversampling

Classifier	Accuracy (%)	Sensitivity	Specificity	F1 score
Naïve Bayes	86.66	0.05	0.95	0.07
C4.5	86.59	0.55	0.92	0.56
Random forest	**86.78**	**0.62**	**0.91**	**0.60**

5 Conclusion and Prospects

The appeal of the proposed work is related to the need to set up a model that forecasts the visitor's shopping intent as soon as the e-commerce website is visited. The advantage behind that is to avoid the risk of abandon implied by each visit on the website. In this scope, the challenging issue consisted in finding out the most appropriate machine learning which could reach this purpose.

Three classification techniques have been investigated to resolve the addressed problem, namely naïve Bayes, C4.5 and random forest. Moreover, oversampling has been carried out to improve the performance and the scalability of each classifier. Based on experimentation and comparison results, we have

proven the efficiency of the random forest classifier as a balanced classifier which is able to fit the requirements of our problem.

As aforementioned, it would be interesting to join a system like that of Sakar et al. [1] to our proposed classifier in order to form a more global system which could be useful for online stores. However, to confirm the effectiveness of a such global system, its exploration in online retailers in real-world and real-time settings should be achieved and its performance should be compared to that of competitive models.

References

1. Sakar, C.O., Polat, S.O., Katircioglu, M., Kastro, Y.: Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Comput. Appl. **31**(10), 6893–6908 (2019). https://doi.org/10.1007/s00521-018-3523-0
2. Mobasher, B., Dai, H., Luo, T., Nakagawa, M.: Discovery and evaluation of aggregate usage profiles for web personalization. Data Min. Knowl. Discov. **6**(1), 61–82 (2002). https://doi.org/10.1023/A:1013232803866
3. Moe, W.W.: Buying, searching, or browsing: differentiating between online shoppers using in-store navigational clickstream. J. Consum. Psychol. **13**(1–2), 29–39 (2003)
4. Poggi, N., Moreno, T., Berral, J.L., Gavaldà, R., Torres, J.: Web customer modeling for automated session prioritization on high traffic sites. In: Conati, C., McCoy, K., Paliouras, G. (eds.) UM 2007. LNCS (LNAI), vol. 4511, pp. 450–454. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73078-1_63
5. Suchacka, G., Skolimowska-Kulig, M., Potempa, A.: Classification of e-customer sessions based on support vector machine. ECMS **15**, 594–600 (2015)
6. Suchacka, G., Skolimowska-Kulig, M., Potempa, A.: A k-nearest neighbors method for classifying user sessions in e-commerce scenario. J. Telecommun. Inf. Technol. **3**(64), 64–69 (2015)
7. Suchacka, G., Chodak, G.: Using association rules to assess purchase probability in online stores. Inf. Syst. e-Bus. Manag. **15**(3), 751–780 (2016). https://doi.org/10.1007/s10257-016-0329-4
8. Budnikas, G.: Computerised recommendations on e-transaction finalisation by means of machine learning. Stat. Transit. **16**(2), 309–322 (2015)
9. Clifton, B.: Advanced Web Metrics with Google Analytics. Wiley, Hoboken (2012)
10. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
11. Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E.: Supervised machine learning: a review of classification techniques. Front. Artif. Intell. Appl. **160**, 3–24 (2007)
12. Baati, K., Hamdani, T.M., Alimi, A.M., Abraham, A.: A new possibilistic classifier for mixed categorical and numerical data based on a bi-module possibilistic estimation and the generalized minimum-based algorithm. J. Intell. Fuzzy Syst. **36**(4), 3513–3523 (2019)
13. Baati, K., Hamdani, T.M., Alimi, A.M., Abraham, A.: A new classifier for categorical data based on a possibilistic estimation and a novel generalized minimum-based algorithm. J. Intell. Fuzzy Syst. **33**(3), 1723–1731 (2017)

14. Baati, K., Hamdani, T.M., Alimi, A.M., Abraham, A.: A modified Naïve Bayes style possibilistic classifier for the diagnosis of lymphatic diseases. In: Abraham, A., Haqiq, A., Alimi, A.M., Mezzour, G., Rokbani, N., Muda, A.K. (eds.) HIS 2016. AISC, vol. 552, pp. 479–488. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52941-7_47
15. Baati, K., Hamdani, T.M., Alimi, A.M., Abraham, A.: A modified Naïve possibilistic classifier for numerical data. In: Madureira, A.M., Abraham, A., Gamboa, D., Novais, P. (eds.) ISDA 2016. AISC, vol. 557, pp. 417–426. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53480-0_41
16. Langley P., Sage S.: Induction of selective Bayesian classifiers. In: Proceedings of 10th Conference on Uncertainty in Artificial Intelligence (UAI-94), pp. 399–406 (1994)
17. Quinlan, J.R.: C4.5: Programs for Machine Learning. Elsevier, Amsterdam (2014)
18. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
19. Subudhi, A., Dash, M., Sabut, S.: Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier. Biocybern. Biomed. Eng. **40**(1), 277–289 (2020)
20. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth and Brooks, Belmont (1984)
21. Baati, K., Hamdani, T.M., Alimi, A.M., Abraham, A.: Decision quality enhancement in minimum-based possibilistic classification for numerical data. In: Abraham, A., Cherukuri, A.K., Madureira, A.M., Muda, A.K. (eds.) SoCPaR 2016. AISC, vol. 614, pp. 634–643. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-60618-7_62
22. Baati, K., Kanoun, S.: Towards a hybrid system for the identification of Arabic and Latin scripts in printed and handwritten natures. In: Madureira, A.M., Abraham, A., Gandhi, N., Varela, M.L. (eds.) HIS 2018. AISC, vol. 923, pp. 294–301. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-14347-3_28
23. Tian, J., Gu, H., Liu, W.: Imbalanced classification using support vector machine ensemble. Neural Comput. Appl. **20**(2), 203–209 (2011). https://doi.org/10.1007/s00521-010-0349-9
24. Ding, A.W., Li, S., Chatterjee, P.: Learning user real-time intent for optimal dynamic web page transformation. Inf. Syst. Res. **26**(2), 339–359 (2015)
25. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)