ISSN 2278-3091
Volume 9, No.3, May - June 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse164932020.pdf
https://doi.org/10.30534/ijatcse/2020/164932020

A Comparative Study of Different Data Mining Algorithms with Different Oversampling Techniques in Predicting Online Shopper Behavior

Ruba Obiedat[1]
[1]King Abdullah II School of Information Technology, The University of Jordan, Jordan, r.obiedat@ju.edu.jo

ABSTRACT

Nowadays buying over the internet has become very popular among online users. This action or attitude has increased due to the facilities and features that e-commerce sites offer to users, such as availability, accessibility, no crowds, easy price comparisons, etc. However, the number of actual buyers is still very low compared to the number of total visitors of these sites. Therefore, this paper will study the behavior of online shoppers to predict whether they will buy a product or not. The study first compares several classification algorithms against each other, then tries to enhance the results using different oversampling techniques. Results show that the RF algorithm achieved the best results and regarding oversampling, the SVMSMOTE exceeded the other methods.

Key words: Click Stream. E-commerce, Online Shopper, SMOTE.

1. INTRODUCTION

The importance and usage of e-commerce have been increased all over the world especially in the B2C world, through online stores. It is expected to continue growing like never before along with the number of online shoppers [1]. Nowadays Online shoppers are expecting higher levels of service and quality with less time, while retailers are trying to understand shoppers' needs and preferences in order to offer them more customized services and products [2]. Unfortunately, the number of real buyers is still very small compared to the number of visitors to these online stores [1]. This raises the need to offer innovative marketing campaigns customized to the shoppers' interests and preferences since regular marketing campaigns have less effect over time on shopping behavior. Moreover, because of the strong competition and business pressure, the necessity to stay competitive no longer optional. The solution for these challenges is to use the right data mining techniques to study shoppers' behavior. This is done through studying their click-stream and historical patterns, in order to recommend a tailored marketing campaign with focused promotions and offers via personalized marketing channels. This leads to higher marketing efficiency, more respondents to the campaign, and an increase in conversion rate as a result [3][4]. This presents the main motivation for this study, together with the difficulty to predict the shoppers' behavior, the essential need to define the most potential customers by retailers to focus on, and the huge amount of data available about the shoppers' behavior and sessions' features and the availability of many data mining tools that can analyze this data [1].

Online websites consist of multiple pages; each page has different links and functionalities. Users interact with the website through different activities such as searching for a certain product, viewing product information, adding a product to the cart, etc. The users' activities and click-stream are stored in user sessions [5]. Each session presents a single visit for a user to the website and may contain various activities. The key activity is the checkout success which indicates that the purchase has been confirmed. Each session is characterized by many features [2]. Different session features and user behavior information are saved on the web server log data as a sequence of HTTP requests sent by the client [5].

Studies prove that the users' navigational behavior presents the main source of information for their preference. Studying users' navigational history can help in predicting their next behavior. Explicit and implicit feedback are one source for a user's online behavior, explicit feedback are easy to collect through feedback forms. However, implicit feedback is the hard one. It can be collected from the users' behavior such as the searches they make, the click-stream they follow and the time they spend on each page and other session information [6].

Today it is easy to access all shoppers' navigational data in each session. . This includes the pages they accessed, the products they viewed and bought, the items they added to the cart, the clicks they made, the time they spent and many others. This is done based on their click-stream and browsing history that is saved on the web server log data [1]. Web prediction applications can solve many important navigation problems as well as improve the training and prediction process. It can save shoppers' time by recommending the next appropriate set of pages, and consequently avoiding visiting unnecessary pages. Additionally, web prediction applications may focus on shopper profiling, as they are categorized according to their interests and preferences, which can be anticipated based on their navigational history [7].

Physical stores employs sales people in order to increase their sales as they can study their customers, know their preferences and offer customized services based on their experience. In the virtual world, website owners try to imitate the sales people’s behavior by investing in the proper data mining techniques in order to analyze the online shoppers’ navigational behavior and session features, so they can expect their needs, offer them more customized recommendation systems, promotions and tailored products. Consequently, this will increase the conversion rate [8]. Retailers apply different online and offline analytical software in order to retain their own customers, keep them satisfied and engaged, as well as, to attract new customers. Web usage mining presents one of the main tools to achieve this [2].

Using the suitable data mining techniques can lead to successful customer profiling and classification, high-quality website navigation design, increased customers satisfaction, better product recommendations customized to customer preferences and many more [9] [1] [10]. Since online stores have a huge amount of information, it becomes a challenge for users to find the most useful one for them. In order to help users get the right information and save time, it is important to build a useful recommendation system customized to each user. Consequently, it becomes necessary to study each user’s behavior in order to offer a better quality of service and build a customized recommendation system [6].

One of the main indicators for the online shoppers’ behavior and purchasing intention is the click-stream data. This presents the navigational path that they have followed with the series of web pages they have requested in a session. Shopper sessions contain data about the user’s navigation history including their visited pages along with the duration for each visit . The click-stream data can be gained from the shoppers’ web server log [1].

Analyzing this data, especially the data with successful purchase experiences, using the appropriate tools may provide in depth insight on the shoppers’ browsing behavior, and help online retailers in many aspects; such as understanding their customers’ needs, recommending the appropriate products, arranging the site services in a way that will increase the number of purchases, maximize the conversion rate, focusing on key customers, using the right marketing techniques, identifying potential customers and converting them into real buyers, highlighting the main reasons for abandonment and taking the right decisions overall[1],[10].

The session is represented by shopper clicks during each visit to the online store. The key sessions are the ones with order confirmation, in which shoppers make a real purchase. Depending on that, the sessions are categorized into buying sessions and non-buying sessions [10]. Studies find that the behavior of the online shopper who actually buys something is different from the one who remains only as a visitor. As aforementioned, this raises the need for data mining tools to analyze customers’ click-stream, history of purchase, as well as, predict online shoppers’ intention to make a purchase [1].

This paper is organized as follows; Section 2 discusses the related work on the topic of online shopper behavior. Section 3 introduces the proposed approach to this work. The experiments and results are given in Section 4. Finally, the conclusion of this work is presented in Section 5.

2. RELATED WORK

Many researchers were interested in the online shopper topic and have tried to analyze it from different perspectives using different tools and algorithms. For example, [7] tried to analyze and study the Markov model and all-Kth Markov model, ARM, two-tier framework, and the modified Markov model in web prediction. The study aimed to increase the scalability, accuracy and prediction time by proposing a modified Markov model to handle the memory requirements, using a new two-tier prediction framework. This framework creates a special classifier, dubbed Example Classifier (EC), in which each example is mapped to the appropriate classifier from a set of available classifiers, and this classifier is then used for prediction. It also studied the Markov model and all-Kth with different N-grams and found out how different N-grams can affect accuracy.

Other research applied the Markov chain prediction model. Such as [6] who applied the Markov chain prediction model in order to predict the user’s next browsing page in a tourism website and built a user preferences matrix by studying the user travel preferences on the website. The study focused on the user implicit feedback behavior in the website. Methods used for web page prefetching were also used to find users’ next Scenic Spot while navigating the website.

Some studies aimed to predict the purchasing intention of online shoppers, [1] studied the behavior of two different shopper groups in a real online book store. The first group; the traditional customers, versus the innovative ones. The study used the association rule mining by applying the Apriori algorithm, in order to predict the probability of making a purchase based on the shoppers’ web server log data, the categories of the products they had visited, and other session features. Results showed differences in the session features affecting the purchasing probability for both groups as well. This study had some limitations, as it is hard to be generalize its findings to other e-commerce sites. Additionally, other session features which had not been discussed in the study may affect the shopper behavior. Also, this research did not address the traffic variations during the time of the day, the day of the week, and the season. Furthermore, other divisions for the customer may be more appropriate for other industries, and most importantly the abandonment of the shopping cart needs more focus and analysis in order to identify the reasons and the behavior pattern that may lead to that abandonment. Some of these limitations are handled by this study.

Moreover, [10] proposed using a back-propagation neural network model to predict shoppers' purchase intention in e-commerce sites using a set of user sessions. This study aimed to anticipate shopper future purchases based on their behavior in the current visit. The data was taken from an online shop that sells books, audio-books, films, in addition to other related items; such as quizzes or short videos. The research studied nine-session features, and multi-layered, back-propagation neural network model was applied. The main limitation was that the number of training samples was low and the number of features were low too.

Some researchers have used clustering techniques in order to classify online shopper sessions into two classes: buying sessions and browsing sessions, based on some session features using the K-nearest neighbor algorithm [5]. The users' session information and features can be used to predict users' probability of completing a purchase order. The data was collected from an online bookstore where the shoppers' historical data stored in the web server log files. The efficiency of the k-NN algorithm was tested with a different number of neighbors. The best results were gained with 11-NN. Furthermore, [2] used Support Vector Machine (SVM) classification based on the users' session features in order to predict the online purchases placed by shoppers. The sessions were also divided into two classes: browsing sessions and buying sessions. The sessions' features were presented with 23- element vector. As for the limitations the classifier proposed by this study may not suit other types of online stores, as the session features and navigational pattern may differ.

On the other hand other researchers focused on finding the key design guidelines necessary to build a high-quality website [9], [11]. These guidelines can be used by website designers to improve website usability, increase number of visitors, and increase customer satisfaction. [11] analyzed user information collected from a virgin olive oil website using descriptive data mining algorithms; such as clustering, association rule, and subgroup discovery for this purpose. Web usage mining was applied within Knowledge Discovery in Database (KDD), in order to enhance the quality of the website design, clustering was used to discover user patterns, while association rules were used to discover association without identifying target class. Finally subgroup discovery was used to find significant relationships according to certain properties such as keywords and source of visit.

In the same context, [9] the main goal was to discuss different data mining techniques used to find the most suitable design for the shopping website in order to enhance the shoppers' navigation experience. The study reviewed some data mining techniques used to classify shoppers based on their navigational behavior and highlighted how shoppers' frequent navigational path affects their behavior. The research discussed the shopping cart as an example of the frequent navigation paths. It highlights the importance of finding the reasons behind the abandonment of the shopping cart and suggests using simple binary classification techniques.

Additionally, [9] reviewed some data mining techniques used to discover the navigation pattern of web pages such as statistical analysis, association rules, clustering, classification and sequential pattern discovery. These patterns then were analyzed to find the most useful one. Web usage mining is also used to discover the frequent browsing items from user history and uses this data to generate more customized recommendations.

Finally [8] who has collected and used the same data of our study, and has analyzed the shoppers' real-time behavior and came up with two modules; the first module aimed to predict the shoppers' intention to buy and the second one had calculated the likelihood of website abandonment. Shoppers' intention to make a purchase was predicted using the shoppers' session information and their historical navigation behavior kept in their web log files. The study applied some data mining techniques such as Random Forest (RF), Support Vector Machines (SVMs), and Multilayer Perceptron (MLP) and results proved that MLP had higher accuracy than others. Another finding was that even though the user's navigational path, click stream and visit history present an essential source of information to predict the purchasing intention, it is important to combine it with session features in order to gain better results.

The second module aimed to predict the user's probability to leave the website without completing the purchasing process, using only sequential clickstream data. Both modules can be used together in order to find out the highest number of potential customers expected to make a purchase, but with the ability to abandon the website before completing the final purchase. Finally, the oversampling method is used to solve the class imbalance problem by increasing the positive class instances in the dataset.

Our work here presents an extension for the work of [8], given that we start by comparing different data mining algorithms as well. Whereas, [8] had only used one common oversampling technique, our work focuses on applying different oversampling techniques to gain the best result. Consequently, we have applied five different oversampling techniques, compared their results, chose the best one, and gained better results in terms of accuracy, F-measure and other evaluation metrics than the other studies.

3. METHODOLOGY

This section outlines, the proposed approach of predicting online shoppers' intention and behavior. First, a detailed data description is presented, then the preprocessing mechanisms applied to the data have been depicted. Additionally, the oversampling techniques and the classification models of the proposed approach are discussed. Finally, a brief description of the evaluation measures has been presented. The proposed methodology followed in this research is shown in Fig. 1 below.

3.1 Data Description

In this work, the dataset describes the user's intention to purchase, whether the process can end with a purchase or not. The data was collected and prepared by [8] and publicly available at the UCI repository. The dataset has 18 features, divided into two types; categorical and numerical features, as shown in Tables 1 and 2. Besides, 12,330 sessions (instances) were extracted, where each session (instance) commonly belong to one user. The collection process was carried outover a one-year period, in order to avoid any tendency from a specific special day, period, user profile or campaign. The study address a binary classification problem with 10,422 (84.5%) out of all the 12,330 sessions did not complete the purchasing process (denote with negative), while only 1908 users success fully purchased something (positive class), this is considered as an imbalanced classes problem that we have handled in our study using different oversampling techniques.

The numerical features shown in Table 1 were extracted from the URL information of the visited pages and updated in real-time, each time the user made an action; such as a switch for another page. Examples of these features are; product related duration (which represents the number of pages and the time spent on each page), Administrative (which is the number of visited pages by the user that related to account management), and Informational (which is the number of visited pages by the user that related to the shopping site's address information and communication).

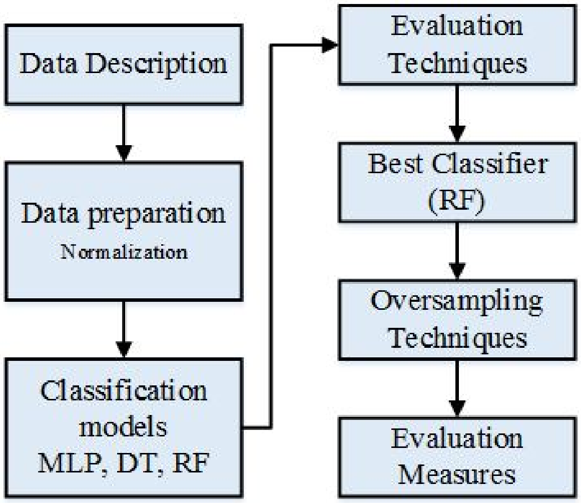

Figure 1: Proposed Methodology.

Table 1 : List of numerical features

Feature name	Min. value	Max. value	SD
Administrative	0	27	3.32
Administrative duration (in seconds)	0	3398	176.70
Informational	0	24	1.26
Informational duration (in seconds)	0	2549	140.64
Product related	0	705	44.45
Product related duration (in seconds)	0	63,973	1912.25
Bounce rate	0	0.2	0.04
Exit rate	0	0.2	0.05
Page value	0	361	18.55
Special day	0	1.0	0.19

Other features such as Exit Rate, Bounce Rate, and Page Value were also extracted from the Google Analytics tool for each visited page (e-commerce site). These values are updated automatically for all visited web pages and are usually stored in the e-commerce site's database. Furthermore, the Special Day feature represents the closeness between the visit time of the website on a particular special day; such as, Valentine's Day or Mother's Day. The value can be determined by taking into account the dynamics of e-commerce; for example, the duration of the order and the delivery.

As for the categorical features shown in Table 2, there are nine presented features that are described; For example, the utilized operating systems and browser while visiting the shopping site, weekends and month features which indicate whether the visit is a weekly or monthly visit. The revenue feature describes if the operation ends with a transaction or not, while the region feature indicates where the session originated from. The remaining features are traffictype and visitor-type, which indicate the source of the traffic and if the user is a new or returning visitor, respectively.

Table 2: List of categorical features

Feature name	Number of categorical values
OperatingSystems	8
Browser	13
Region	9
TrafficType	20
VisitorType	3
Weekend	2
Month	12
Revenue	2

3.2 Data Preparation

In this subsection, the data goes through several processes to be ready for the final step, in which it is examined by the classifiers. One of these processes is the normalization of the data, which is a technique that occurs as a part of the data preparation process for the classification models. The normalization can be applied by modifying the values of the numeric features to a specific scale[4]. In this study, the normalization is performed by using the min-max criteria, where the values will be changed into the range [0,1] as shown in Eq. 1.

$$B = \frac{A - min_A}{max_A - min_A}(max_B - min_B) + min_B \quad (1)$$

where the new scaled value is denoted as B, and the value needed to be scaled is A. The lower and upper bounds of the previous interval is referred to as minA and maxA, respectively. Moreover, minB and maxB are the lower and upper bounds of the novel interval.

3.3 Classification Models

This subsection briefly reviews the different classification algorithms which have been applied in this study. Three commonly used techniques were chosen:

Multilayer Perceptron (MLP) Multilayer Perceptron (MLP) is considered one of the most popular kinds of artificial neural network algorithms. The MLP can be defined as a complex modeling method that was inspired by human neurons. It can also be described as a processing system composed of several layers; an input layer, a hidden layer, and an output layer, respectively [12]. The MLP can be referred to with another concept in literature, namely feedforward artificial neural network. The basic element of the MLP or the feedforward neural network is the perceptron, where each one of the perceptrons presented with multiple inputs. Furthermore, these inputs shift into the summation function and activation function in order to produce an output for this perceptron. The activation function has different types of functions, however, the most used one is the sigmoid function, where it is responsible for scaling the values to be between 0 and 1 [13].

Decision Tree (DT) The Decision Tree (DT) is composed of directed edges and nodes, and it is shaped as a flowchart-like tree structure that divides the data into different segments[14]. These nodes have three types, the first one is the root where it is the node from the top and does not have any incoming edges. The second type is usually placed in the middle of the tree and is known as the internal node. While the last node with no outgoing edges, is placed at the end of the tree and is called decision node or the leaf. The internal node divides the attribute values of the dataset, then assigns the values to the last node (leaf), where each path forms a decision [15]. One of the common algorithms is based on the DT method, and it is C4.5 [16]. The C4.5 algorithm uses an enhanced tree pruning for decreasing misclassification errors. The algorithm utilizes the information gain technique in order to produce the trees, then calculates the attribute values and orders them based on the information gain values, root, internal node, and leaves[16].

Random Forest (RF) Recently, the Random Forest (RF) is considered to be one of the best classification models in literature [17]. It is known as an ensemble method, where the base classifier is the DT algorithm. The RF generates a tree by a random subset from the training data, then a random selection is utilized on the attributes in order to determine the split [18]. Therefore, the RF generates uncorrelated and independent trees that depend on random inputs. The evaluation criteria for the produced decision trees occur throughout the Out of Bag criteria, where the non-selected (remaining) data are used. Subsequently, RF applied the majority voting to determine the final decision, where it runs efficiently and rapidly in most cases without overfitting problems [19].

3.4 Oversampling Techniques

Five different oversampling techniques were applied on the online shoppers' dataset in order to overcome the imbalanced data class problem and improve the classifiers performance.

SMOTE Synthetic Minority Over-sampling Technique (SMOTE) is the most common oversampling mechanism used on the unbalanced class distribution [20], where it generates new points of data instead of replicating the points itself. SMOTE depends on the nearest neighbor technique to operate, in order to find the n nearest neighbors for the data points from the minority class. It selects the n randomly from the closest neighbors based on the oversampling value while using the Euclidean Distance method. Thus, the new synthetic point will be chosen as a point from the path between the original data and the selected neighbor. As a result, a balanced distribution will occur between the minority and majority classes [21].

ADASYN It is considered as an improved version of SMOTE with almost the same oversampling techniques to handle the imbalanced data classes. The adaptive synthetic (ADASYN) operates by creating a sample and then adding small values to the points randomly [22]. In other words, the sample will have little variance values more than the original data, and this will make the points more realistic.

SMOTE-Borderline Borderline-SMOTE (BSM) is another extension of the standard SMOTE proposed by [23]. However, unlike SMOTE the BSM generates synthetic instances close to the boundaries. This will cause for the borderline instances to be misclassified more than the far ones. Consequently, this improves the classifier performance [23].

SVM-SMOTE It's an oversampling technique that expanded form the SMOTE mechanism. The SVM-SMOTE is applied to overcome the imbalance problems in data. SVM-SMOTE synthetically produce new instances from the minority class of their nearest neighbors. Furthermore, the minority class instances are generated near the borderlines with the help of SVM to initiate boundaries between each class [24].

SMOTE-NC Synthetic Minority Oversampling Technique -Nominal Continuous (SMOTE-NC) is applied to solve unbalanced class distribution [20]. It generates synthetic instances for quantitative and categorical attributes, where the SMOTE-NC generates the new sample to be more specific for the categorical attributes by selecting the most frequent attribute of the nearest neighbors.

3.5 Evaluation Measures

To evaluate the results of our model, a variety of evaluation metrics were used. These measures are computed using the confusion matrix illustrated in Table 3. which is considered as the most common method used to evaluate the performance of the classification model[25], and they are Accuracy, Recall, Precision and F1 Score conformed to the formula as used in many studies such as[26][27].

Accuracy is the proportion of instances that were correctly classified by the model. Its formula is defined by Eq. 2.

$$Accuracy = \frac{TP+TN}{TR+TN+FP+FN} \quad (2)$$

Recall is the proportion of relevant results that were correctly classified by the model. There are two types of recall; sensitivity and specificity. They are defined by Eq. 4 and Eq. 3.

$$Specificity = \frac{TN}{TN+FP} (3)$$

$$Sensitivity = \frac{TP}{TP+FN} (4)$$

Precision is the ratio of relevant retrieved results to the number of retrieved results. Eq. 5 defines its formula.

$$Precision = \frac{TP}{TP+FP} (5)$$

F1 Score is the harmonic mean of Precision and Recall, and it is defines by Eq. 6.

$$F1Score = \frac{2 \times Precision \times Recall}{Precision \times Recall} (6)$$

Table 3: Confusion Matrix

	Actual buyers	Actual non-buyers	Total instances
Predicted buyers	TP	FN	P
Predicted non-buyers	FN	TN	N

4. EXPERIMENTS AND RESULTS

4.1 Classification Results

In this section, three classification algorithms have been applied and compared to each other using a variety of measurements. Accuracy, precision, recall and F measure of Random Forest, Multi-Layer Perceptron and Decision Tree have been measured. And as illustrated in Table 4, Random Forest had the best performance among the other classifiers; it achieved 89.1%, 60.7% and 69% for accuracy, F1 score and precision, respectively. While the second best classifier obtained by the Decision Tree achieved 85.0% in accuracy manner. As for the other measures, the classifier achieved 53.5% in F1 score and 53.0% in precision. On the other hand, the MLP did not compete with the other classifiers and achieved low results in all measures. Although the Random Forest has outperformed the other classifiers, some of the values were not very interesting; like the recall for the positive class which was 55.1%.

As mentioned earlier, Random Forest algorithm obtained the best results. Hence, Random Forest was chosen for the next step in order to apply the different oversampling techniques to enhance the previous results, then select the oversampling method with the best performance measures, therefore a new confusion matrix has been carried out for the classifier as shown in Table 5.

Table 4: Results for Random Forest, MLP and Decision Tree.

Measure	**RF**	**MLP**	**DT**
Accuracy	0.891	0.155	0.850
F1-Positive	0.607	0.268	0.535
F1- Negative	0.938	0	0.908
Precision- Positive	0.690	0.154	0.530
Precision- Negative	0.919	NA	0.912
Recall- Positive	0.551	1.000	0.527
Recall- Negative	0.953	0.000	0.907

Table 5: Confusion Matrix for Random Forest.

	Predicted			
Actual		P	N	Total
	P	1051	857	1908
	N	489	9933	10422
	Total	1540	10790	

4.2 Oversampling Results

Different oversampling techniques were used in order to enhance previous results. The Random Forest model was chosen as the classifier when comparing the results, as it achieved the best results. The five applied oversampling techniques were SMOTE, ADAYSIN, BorderlineSmote, SVMSmote and SMOTENC. The performance of the oversampling techniques was studied using a number of oversampling percentages to select the best percentage that achieves the highest F measure. Fig. 2 summarizes these results and shows that the performance was the best at 446.26% percentage. Figure 3 shows a comparison between SMOTE, ADAYSIN, BorderlineSmote, SVMSmote and SMOTENC, based on F measure for the negative class. As we can see, SVMSMOTE achieved the best results. Moreover, the regular SMOTE obtained very competitive results in the SVMSMOTE with almost 0.923. As for the positive class, the SVMSmote also outperformed the other methods with 0.915 and followed by the SMOTENC with nearly 0.914. These results considered quite high comparing with other studies results mentioned previously.

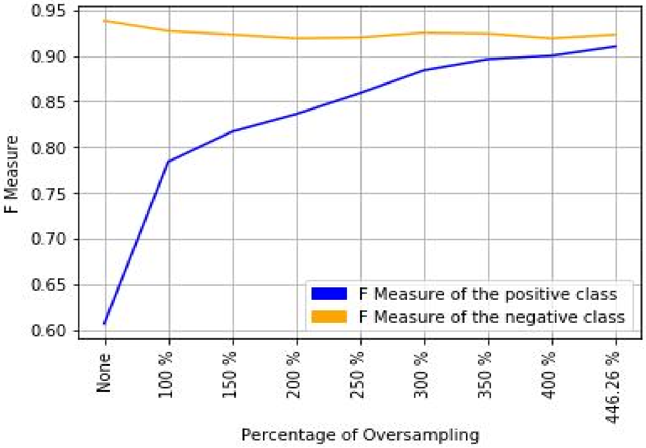

Figure 2: F Measure Scores for SMOTEusing several oversampling percentages.

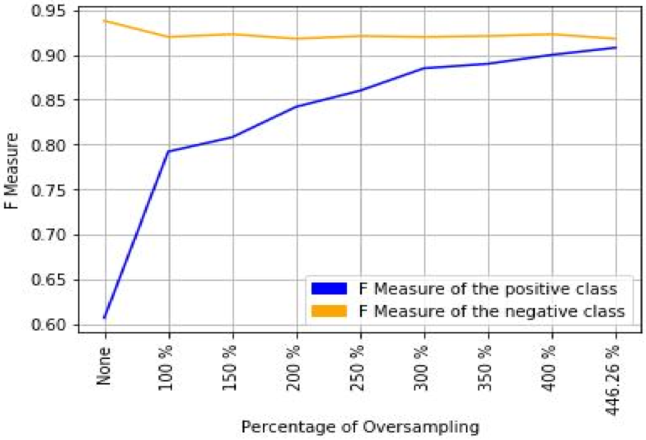

Figure 3: F Measure Scores for ADAYSINusing several oversampling percentages.

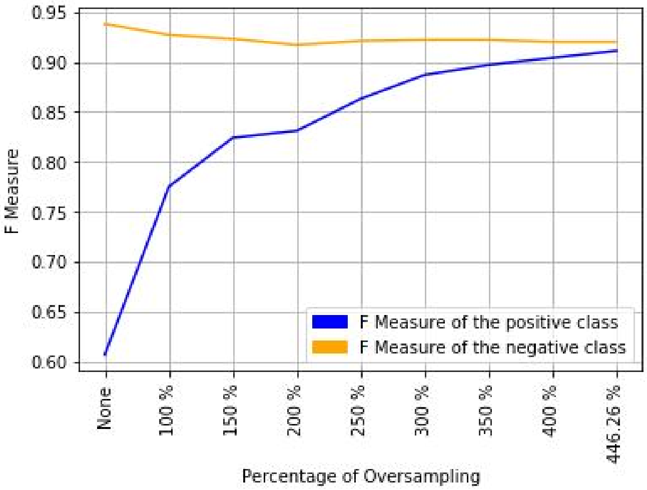

Figure 4: F Measure Scores for BorderlineSmoteusing several oversampling percentages.

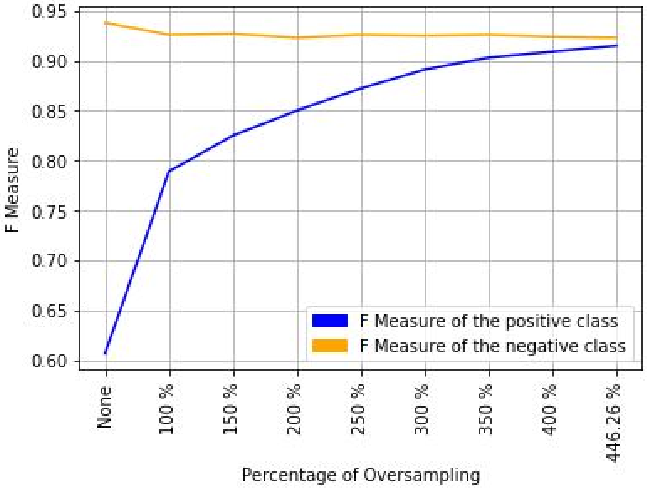

Figure 5: F Measure Scores for SVMSMOTEusing several oversampling percentages.

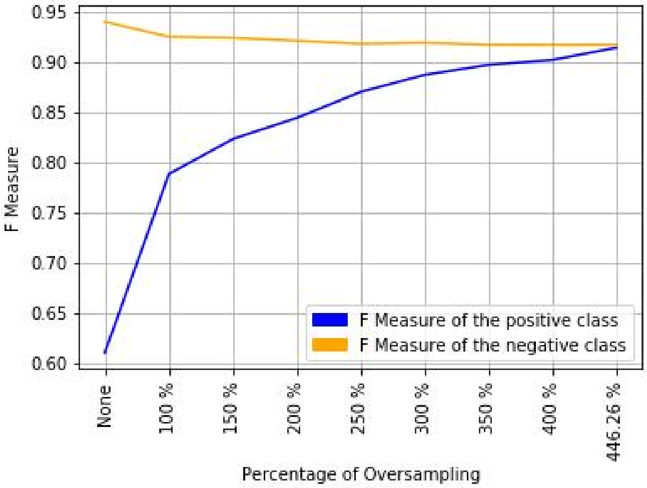

Figure 6: F Measure Scores for SMOTENCusing several oversampling percentages.

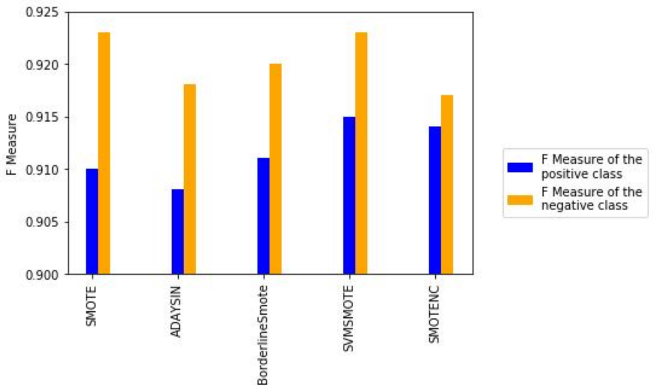

Figure 7: Comparison between Oversampling Techniques based on F Measure; and after choosing the best oversampling percentage, which is 446.26 %.

5. CONCLUSION

In the past few years the online hopping had increased dramatically and became a trend topic, unfortunately the conversion rate has not increased at the same level, many researcher tried to study the reasons behind that and find solutions regarding this issue. In this paper, a comparison between the three classification models was carried out in order to study the behavior and intention of online shoppers. These models were Random Forest, MLP and Decision Tree. The Random Forest outperformed the other classifiers in accuracy, F1-score, and precision, with 89.1%, 60.7%, and 69%, respectively. However, not all results were satisfying, particularly the recall measure. Thus, oversampling techniques have been employed in order to enhance such results. In this manner, we have applied five of the the best oversampling techniques which are; SMOTE, ADASYN, SMOTE-Borderline, SVM-SMOTE and SMOTE-NC results shows that the best oversampling percentage that achieves the highest performance in term of F measure was 446.26% percentage. Moreover, comparing the five oversampling techniques the best results accomplished by the SVMSMOTE for the negative class with more that 92.3%, while the SMOTENC has the best results for the positive class with 91.5% which considered relatively high.

REFERENCES

1. G. Suchacka and G. Chodak. **Using association rules to assess purchase probability in online stores**, *Information Systems and e-Business Management*, vol. 15, no. 3, pp. 751–780, 2017. https://doi.org/10.1007/s10257-016-0329-4
2. G. Suchacka, M. Skolimowska-Kulig, and A. Potempa.**Classification of e-customer sessions based on support vector machine,***ECMS*, vol. 15, pp. 594–600, 2015. https://doi.org/10.7148/2015-0594
3. S. Moro, R. Laureano, and P. Cortez, **Using data mining for bank direct marketing: An application of the crisp-dm methodology,***in Proceedings of European Simulation and Modelling Conference-ESM*, EUROSIS-ETI, 2011 pp. 117–121.
4. E. Turban, R. Sharda, and D. Delen, **Business intelligence and analytics: systems for decision support.***Pearson Higher Ed*, 2014.
5. G. Suchacka, M. Skolimowska-Kulig, and A. Potempa, **A k-nearest neighbors method for classifying user sessions in e-commerce scenario,***Journal of Telecommunications and Information Technology*, 2015.
6. Y. Shi, Y. Wen, Z. Fan, and Y. Miao, **Predicting the next scenic spot a user will browse on a tourism website based on markov prediction model,** *in 2013 IEEE 25th International Conference on Tools with Artificial Intelligence*, , IEEE, 2013, pp. 195–200.
7. M. A. Awad and I. Khalil, **Prediction of user's web-browsing behavior: Application of markov model**, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 1131–1142, 2012. https://doi.org/10.1109/TSMCB.2012.2187441
8. C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, **Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and lstm recurrent neural networks**, *Neural Computing and Applications*, pp. 1–16, 2018.
9. W. L. YEUNG, **A review of data mining techniques for research in online shopping behaviour through frequent navigation paths**, 2016.
10. G. Suchacka and S. Stemplewski, **Application of neural network to predict purchases in online store**, *in Information Systems Architecture and Technology: Proceedings of 37th International Conference on Information Systems Architecture and Technology–ISAT 2016–Part IV*, Springer, 2017, pp. 221–231. https://doi.org/10.1007/978-3-319-46592-0_19
11. C. J. Carmona, S. Ram´ırez-Gallego, F. Torres, E. Bernal, M. J. delJes´us, and S. Garc´ıa, **Web usage mining to improve the design of an e-commerce website: Orolivesur. com**, *Expert Systems with Applications*, vol. 39, no. 12, pp. 11243– 11249, 2012.
12. C. M. Bishop et al., **Neural networks for pattern recognition.** *Oxford university press*, 1995. https://doi.org/10.1201/9781420050646.ptb6
13. M. T. C. Olmedo, M. Paegelow, J.-F. Mas, and F. Escobar, **Geomatic approaches for modeling land change scenarios**. *Springer*, 2018.
14. J. Han, J. Pei, and M. Kamber, **Data mining: concepts and techniques.** Elsevier, 2011.
15. H. Dahan, S. Cohen, L. Rokach, and O. Maimon, **Proactive data mining with decision trees.***Springer Science & Business Media*, 2014.

16. J. R. Quinlan, **Induction of decision trees,***Machine learning*, vol. 1, no. 1, pp. 81– 106, 1986.
17. M. Fern´andez-Delgado, E. Cernadas, S. Barro, and D. Amorim, **Do we need hundreds of classifiers to solve real world classification problems?,***The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.
18. R. Malhotra and R. Jangra, **Prediction & assessment of change prone classes using statistical & machine learning techniques.,***Journal of Information Processing Systems*, vol. 13, no. 4, 2017.
19. L. Guo, N. Chehata, C. Mallet, and S. Boukir, **Relevance of airborne lidar and multispectral image data for urban scene classification using random forests**, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 1, pp. 56–66, 2011. https://doi.org/10.1016/j.isprsjprs.2010.08.007
20. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, **Smote: synthetic minority over-sampling technique**, *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
21. S. Kotsiantis, D. Kanellopoulos, P. Pintelas, et al., **Handling imbalanced datasets: A review**, *GESTS International Transactions on Computer Science and Engineering*, vol. 30, no. 1, pp. 25–36, 2006.
22. H. He, Y. Bai, E. A. Garcia, and S. Li, **Adasyn: Adaptive synthetic sampling approach for imbalanced learning**, *in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, IEEE, 2008, pp. 1322–1328.
23. H. Han, W.-Y. Wang, and B.-H. Mao, **Borderline-smote: a new over-sampling method in imbalanced data sets learning**, *in International conference on intelligent computing*, Springer, 2005, pp. 878–887.
24. B. Fergani et al., **Comparing hmm, lda, svm and smote-svm algorithms in classifying human activities**, *in Proceedings of the Mediterranean Conference on Information & Communication Technologies*, Springer, 2016, pp. 639–644. https://doi.org/10.1007/978-3-319-30298-0_70
25. Sdw T. Mauritsius , A. S. Braza , Fransisca, **Bank Marketing Data Mining using CRISP-DM Approach,** *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)* Vol. 8, No. 5, 2019 https://doi.org/10.30534/ijatcse/2019/71852019
26. A. Z. Maghuyop, **A Response Assessment on the Implementation of Senior High School TVL Track through Data Mining Technique,** *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE),* Vol. 8, No.6, 2019 https://doi.org/10.30534/ijatcse/2019/06862019
27. N. S. Buot, **Multiple Intelligences and Reading Comprehension of Senior High School Students: A Response Evaluation through Educational Data Mining Technique**, *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE),* Vol. 8, No.6, 2019 https://doi.org/10.30534/ijatcse/2019/30862019